User-Centric Analytics for Expert Crowds

by

Iuliia Chepurna

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Applied Science

in

Electrical and Computer Engineering

University of Ontario Institute of Technology

Supervisor: Dr. Masoud Makrehchi

September 2015

Abstract

Pervasiveness and permissiveness of social media made content verification an integral part of any analysis involving online generated data. The most natural and convenient way to assess content trustworthiness and validity is to examine it from the perspective of a user that authored it. With more studies switching their strategies of data collection from event-oriented to user-focused, it is crucial to outline and address the challenges pertaining to this approach. We propose a *user-centric analytics* paradigm as one of the solutions to this problem. It is casted as a three-tier framework, consisting of detection of topical experts, extraction and interpolation of their opinions and utilization of the latter in *social filtering*. The first is concerned with automatic identification of user's topical attribution on Twitter: while the platform is so popular among professionals, it does not support an explicit mechanism for community membership. We present three models exploiting semantic signature of a group and examine their performance on a case study of Twitter investment community. Interpolation of missing opinions is intended to handle mass amounts of periods with no activity peculiar to user streams. We introduce a number of community-based models exploiting user's historical activity, content and opinions of his immediate network, and also verify their feasibility to serve as an initialization scheme in low-rank matrix approximation. We also analyze how predictability changes from user to user and build a model based on his characteristics to assess this value beforehand. Finally, we

present a concept of *social filtering*—an approach with objective to exploit plethora of available historical data for prediction of social trends. In contrast to collaborative filtering, it does not solicit explicit recommendations from users and operates on raw data. It is designed to automatically select "expert" users—individuals whose content is the most reflective of a target central to the application—and transform their posts into predictive signals.

Acknowledgements

I would like to express gratitude to my supervisor, Dr. Masoud Makrehchi, for being good friend and advisor both inside and outside of academia. Thank you so much for sharing your knowledge and always supporting me no matter what. I will never forget your entrepreneurial spirit, bursts of unique ideas, and thought-provoking discussions. Working with you was a great pleasure and unforgettable experience, and it is an honor to be one of your first students.

Kyr Shegeda, love of my life, you have always been my hero, inspiration, and perfect role model. You constantly remind me that hard work and dedication can make the most incredible things happen. Thank you for your contagious optimism, confidence and ever-seeking mind that helped me to persevere during the hardest times. I cannot express how much grateful I am for your selfless support and infinite faith in me.

My sincere appreciation goes to Bahar Aghababaei, a friend, colleague and sister that I have never had. Thank you for being a part of incredible memories, for always helping even when I did not ask, for listening, cheering and understanding. I owe you so much. Also many thanks to my old friends and the ones I have met along the way, without you this journey would have never been possible.

Finally, I would like to thank my parents for their unconditional love and sacrifices they have made to provide me with the better life. My deepest love and respect go

to my grandma who dreamt of this moment but did not have a chance to share my joy.

To my grandmother

Contents

Abstract **i**

Acknowledgements **iii**

Contents **vi**

List of Figures **viii**

List of Tables **xi**

1 User-Centric Social Analytics **1**
1.1 Motivation 1
1.2 Our Approach 3

2 Topical Experts Detection **9**
2.1 Introduction 9
2.2 Previous Work 11
2.3 Our Approach 16
2.3.1 Profile-Based 20
2.3.2 Behavior-Based 21
2.3.3 Domain-Specific Filter 22
2.4 Experimental Results 24
2.4.1 Dataset Description 24
2.4.2 Preprocessing 25
2.4.3 Experimental Scenarios 25
2.4.4 Discussion 28
2.5 Conclusion 32

3 Interpolation of Missing Opinions **35**
3.1 Introduction 35
3.2 Previous Work 38
3.3 Our Approach 42
3.3.1 Selected domain 42
3.3.2 Bull-Bear Classifier 44

3.3.3 Community-Based Inference . . . 46
3.3.4 Individual Predictability . . . 48
3.3.5 Low-Rank Matrix Approximation . . . 49
3.4 Experimental Results . . . 52
3.4.1 Dataset Description . . . 52
3.4.2 Bull-Bear Classifier . . . 53
3.4.3 Community-Based Inference . . . 54
3.4.4 Individual Predictability . . . 57
3.4.5 Low-Rank Matrix Approximation . . . 60
3.5 Conclusion . . . 61

4 Social Filtering **63**
4.1 Introduction . . . 63
4.2 Previous Work . . . 66
4.3 Our Approach . . . 67
4.4 Experimental Results . . . 68
4.4.1 Dataset Description . . . 69
4.4.2 User Representation . . . 70
4.4.3 Experts Selection . . . 73
4.4.4 Prediction *vs.* Reporting . . . 77
4.4.5 Dependency on Historical Data . . . 78
4.4.6 Predictability by Type . . . 79
4.5 Conclusion . . . 82

5 Summary of Contributions and Future Work **86**
5.1 Contributions . . . 86
5.2 Future Work . . . 89

Bibliography **92**

List of Figures

2.1 Example of Stocktwits posts. Real usernames and avatars are replaced. 20
2.2 Cumulative distribution function of a fraction of relevant tweets in timelines of (a) *golden* Stocktwits and (b) *target* Twitter datasets as determined by *domain-specific filter*. μ and $\mu_{1/2}$ denote mean and median of the distributions. 23
2.3 Experimental scenario 1.1. Classifier is fitted on *golden* Stocktwits and *white noise* datasets. It is assessed using 10-fold cross-validation. . . 26
2.4 Experimental scenario 1.2. Classifier is fitted on *golden* Stocktwits and *white noise* datasets. It is tested on *target* Twitter and *white noise* dataset. It is a target scenario in which model learns from automatically selected data to determine relevance of target users. 27
2.5 Comparison of the results yielded by proposed models. Models are fitted on *golden* Stocktwits and *white noise* datasets. In scenario 1.1 they are tested using 10-fold cross-validation, in scenario 1.2—*target* Twitter and *white noise* datasets. 31
2.6 Distribution of content generation in *target* Twitter dataset 31
2.7 Dependency of accuracy on the size of available timelines when predicting target users from Twitter dataset 32

3.1 Rastergram of daily opinions shared by 100 most active users of Stocktwits Apple community captured during 730 days of observation. Black vertical bar indicates that a user had at least one post on that day. Ideal data would resemble a black square. 36
3.2 Example of Stocktwits posts with *Bullish* and *Bearish* annotations. Real usernames and avatars are replaced. 43
3.3 Opinion-level performance yielded by sentiment- and content-based models, majority vote and historical activity baselines. F_1^- and F_1^+ stand for F-measure of negative and positive classes respectively. . . . 56
3.4 ROC curve of opinion-level performance when inferring *bullish* (a) and *bearish* (b) opinions. *Random choice* also resembles ROC for *always bullish* and *always bearish* baselines. 57

3.5 Probability density function of individual predictability yielded by *sentiment-based* model. Although F-measure averaged over 100 users is equal to 0.57 only, almost 40% of users exhibited individual predictability above this value. 58
3.6 Relationship between variance in individual F-measure across k folds and overall user's predictability. Pearson's $r = 0.54$. 58
3.7 F-measure yielded by different feature sets for user predictability model 59
3.8 Matrix disparity before (a) and after (b) low-rank approximation (r=2) for different amount of data being held out 60
3.9 Improvement in matrix disparity when comparing original matrix with initialized and approximated matrices accordingly 61

4.1 Collaborative problem solving with regards to content generation. By *personal motivation* we mean one's enthusiasm and desire to publish some content that is not fostered by external factors. 65
4.2 Daily aggregated indicents. Spikes were observed during statutory holidays. 69
4.3 Daily tweet volume. Peaks are labeled with corresponding topics trending in our set during those days. 71
4.4 Performance of user representation for varying lag. *Negative*, *sad* and *anxious* affects as well as fractions of *obscene* and *death*-related words exhibit the same trend, hence we report only *negative affect*. 72
4.5 Daily number of incidents with corresponding positive affect aggregated across all users 72
4.6 User activity in terms of days with published content and total number of posts. (a) Daily number of unique active users. Total number of users between July 1, 2010 and November 30, 2013 is 2753. On average only 15% of all users contribute to daily content. (b) Distribution of posts between users. Log-log scale is used. 74
4.7 (a-c) F-measure obtained during each phase of three-step filtering. Rows correspond to single experiments. Lighter colors stand for higher values. (d) Comparison of the best results: (i) before filtering (all 2753 users are kept), (ii) for activity-based, (iii) relevance-based filters and (iv) ensemble of users. 75
4.8 Predictability observed for different leads and lags. Areas corresponding to *reporting* and *predicting* behavior lie to the left and right of $l = 0$ respectively. Best predictability occurs on $l = 7$ and $l = 14$. 77
4.9 Experiment scenarios. Test set is equal to 20% of the dataset and is fixed between April 1, 2013 and November 30, 2013. Scenario 1 (a) uses training set starting in July 1, 2010 and progressing towards test set, while for scenario 2 (b) starting point is a beginning of a test set and it increases in retrospective direction. 79

4.10 Model performance for scenarios: (a) classifier deterioration and (b) dependency on historical data. Top axes identify the size of corresponding training set. 80
4.11 F-measure and AUC of the model for crime types which yielded the best performance. Lags (in days) are specified on x-axis. $l = 0$ stands for the same-day prediction. 81

List of Tables

2.1 Part of a sample timeline of a stock market expert from *target* Twitter dataset 18
2.2 Part of a sample timeline from *White noise* dataset 19
2.3 Statistics on the size of a user timeline for all used datasets 25
2.4 Comparison of the results yielded by proposed models. F_1^+ and F_1^- stand for F-measure of positive and negative classes respectively. PB denotes *profile-based* model, BB—*behavior-based* model and DSF—*domain-specific filter*. In scenarios 1.x model is trained on golden set and *white noise*, in scenarios 2.x—on target set and *white noise*. Scenarios x.2 denote cross-validation, and x.1—testing on the target and golden set correspondingly. 29

3.1 Sample Stocktwits posts 45
3.2 List of activity, content and network features chosen to represent user profile 50
3.3 Statistics on the timelines of selected users in the pool 53
3.4 Performance of Bull-Bear classifier yielded by lexicon-based and supervised models 54
3.5 User-level F-measure yielded by classifiers employed in *sentiment-* and *content-based* models 55
3.6 User-level performance yielded by sentiment- and content-based models, majority vote and historical activity baselines. F_1^- and F_1^+ stand for F-measure of negative and positive classes respectively. 55
3.7 Importance of individual features (content and activity only) as defined by Random Forest classifier 59

4.1 Number of incidents reported per each of 29 categories 84
4.2 Statistics on user activity during period of observation. Total number of users: 2753. Total number of days: 1249. Users active more than average number of days: 1035 (83%); users with number of posts above average: 892 (32%). 85

Chapter 1

User-Centric Social Analytics

1.1 Motivation

Our democracy is the greatest example of "wisdom of crowds" in action. Given the fact that everyone has a chance to vote, such system is deemed to select the government that is the best fit for the society. This principle is applicable to virtually anything with a single caveat: pool of individuals participating in the discourse as well as their opinions have to be diverse [87]. Although beautifully intuitive and supported by mass evidence, exercising this idea in scientific studies has been somewhat limited by impossibility of, first, surveying all members of a group, and, second, ensuring that respondents have provided honest and unbiased information.

Advent of Web 2.0, in turn, has transformed the core idea of free speech and self-expression and released aforementioned burden completely. Users are now invited to produce high quality content, actively participate in open discussions, share unique professional knowledge by answering questions and contributing to online encyclopedias. Notions of semi-anonymity and social recognition further foster users to manifest their deepest beliefs, concerns and ideas. Unprecedentedly vast amounts of

publicly available information made it possible to tap into collective intelligence and successfully use it for early earthquake alerts [28,35,79], tracking flu [24,30,56,81], predicting outcomes of elections [63,90] and sport games [51,82,91], discovering trending topics [17, 69, 72], and for many other applications [108].

Most of these studies use the content aggregated with respect to event of interest, which has a number of drawbacks. First, malicious posts produced by spammer accounts or cluster of users involved in online puppetry can hardly be identified on their own, and thus can introduce noise and misrepresentation of the actual situation. Second, some users are naturally more active than others, which can lead to their information dominating the dataset, no matter if it actually has more value or not. Finally, some people are prone to be overly pessimistic (or optimistic) in expressing themselves with regards to arbitrary issues. Hence if constantly extreme opinions are not discounted, this can result in distorted view of a problem. Whereas when such content is considered in a context of unique individual, it facilitates the filtering process by easily damping the importance of data produced by a misleading user or completely discarding it. In fact, it was shown for Twitter that the content generated by a pool of users deemed as "experts" in different domains outperformed Twitter native random stream in a number of categories, including popularity, credibility, topical diversity, and variance of expressed opinions [109].

Howbeit, more and more studies in social computing and behavioral modeling now adopt this approach to data processing. Starting with the works where user-level aggregation seems more natural, such as analysis of web search behavior with respect to pregnancy-related concerns [38] and prostate cancer [76], prediction of prospective depression [33], level of users' well-being [80] and satisfaction with life [27]; followed by detection of antisocial behavior [21], online sybils [94] and content credibility [4, 43, 53], where such strategy is less intuitive. Moreover, spread of information in

social network can be predicted better if modeled in a context of participants and their characteristics rather than the sole content [20, 26]. Interestingly, not only applications, for which preserving one's personality is crucial, can benefit from user-attributed timelines—such data can be successfully incorporated in various prediction frameworks to enhance their performance. For instance, tweets were shown to be informative of unemployment rate [64] and stock market movements [60]. For the latter authors have proven that a smaller subset of "expert" users yields even better results than the whole pool [9, 46, 60, 95].

Clearly, there is a need in a unified framework that encompasses appropriate content sampling and subsequent filtering, its enrichment, further utilization and challenges pertaining to each phase. Current work is an attempt to solve this problem from a user-centric perspective. Although not purely abstract, proposed approach is general enough to bring the common issues to the spotlight and demonstrate some viable solutions. We discuss research questions and our contributions in the next section.

1.2 Our Approach

With the previous discussion in mind, we would like to introduce a user-centric paradigm universal enough to be capable of sampling content from social media that is relevant to an arbitrary task, no matter if its goal is sole user modeling, opinion inference, prediction of social trends or any other. As mentioned before, even if an outer model can tolerate noise to some extent, in case of overabundance it may produce erroneous insights and result in plausible conclusions. To account for that, we specifically frame the content processing and aggregation task in terms of user timelines.

Our ultimate goal is then to harvest this cleaned data to select the information that is the most representative of analysis target. We pose this problem as a *social filtering*. Similarly to collaborative filtering, it operates on users' opinions and perceptions, however, the main difference lies in the fact that such "recommendations" have been never requested. Hence, *social filtering* leverages only candid posts that were shared voluntarily, this way preserving authenticity of opinions and decreasing the bias. This approach takes an advantage of massive amounts of already available historical data. Its objective is a detection of a set of "expert" users, those which are the most informative of a target problem, and transformation of their content into a target space.

We cast *user-centric analytics* as a three-tier framework. First, a diverse pool of "expert" individuals needs to be identified. Then their content has to be processed and augmented to mitigate the effect of missing data. And, finally, the *social filtering* must be carried out to determine the most relevant content and translate it into predictive signals. Although each task is equally important and interesting on its own, we prefer to study them under umbrella of user-focused paradigm. We discuss the hypotheses and research questions associated with each of these phases below.

Topical experts detection. Inspired by the "wisdom of crowds" [87] idea, we aim to automatically find a group of distinct users actively sharing their insights on a topic central to application. While the task can be quite trivial if conducted on platforms supporting explicit enrolment to specific groups, medium of our focus—Twitter—does not provide such functionality. Thus it becomes crucial to identify such communities of practitioners. Applications of this component are not limited to generating diverse pools—it can be successfully used for fine-grained professional follower recommendation, retrieving emerging trends in particular fields, business intel-

ligence and many other. Normally there are two concepts—expertise and topicality—associated with skilled individuals. While the first is quite subjective and often hard to assess, here, we prefer to focus on the second, considering expertise as an attribute introducing diversity. Thus, the task is narrowed down to automatic detection of individuals belonging to a topical group of interest. We cast it as a supervised learning problem judging whether a particular person belongs to a community of interest or not based on his public timeline. We introduce hypotheses and research questions below (henceforth **H** and **RQ** respectively).

H1 Members of specific community (*e.g.* computer scientists, cardiologists, politicians, *etc.*) share same lexicon, style of writing and use jargon developed within a group. Hence, such *semantic signature* can be used to determine one's topical attribution.

RQ1.1 Is user's static content (*i.e.* aggregated timeline) representative of his professional interests?

RQ1.2 Does dynamic representation of user's content provide additional context to understanding his topicality? Does it reflect the changes in user's behavior (*e.g.* consider scenario where a user has recently joined the group and has not yet shown enough activity to be related to a community)?

RQ1.3 Is naïve lexicon-based approach able to yield satisfactory performance, or supervised learning is essential to proper identification?

However, the main challenge lies in the fact that there is no training data readily available. While the positive examples of community language can be taken from virtually any source (and it does not have to originate from social media), samples of negative class need to be defined.

H2 ("class bias") Since the fraction of conversations related to particular narrow group on Twitter is infinitesimal, a randomly sampled post has a high probability to be irrelevant to community of interest.

RQ2.1 Is training data generation approach introduced by *class bias* hypothesis sufficient for a particular task?

We analyze these hypotheses and answer research questions in Chapter 2.

Interpolating missing opinions. After *experts* are identified, their timelines need to be refined, since the absent data is imminent. That is no matter how much active chosen users are, for a reasonable choice of time aggregation, there would always be activity gaps. Similarly to noise, while decent models are expected to provide some level of robustness, it is hard to perform in a setting where at least half of the data is missing. Thus, we posit that users' missing opinions can be interpolated based on their historical activity and immediate network.

H3 User's opinion is a function of (i) his previous activities, (ii) opinions of his friends and (iii) a discourse of immediate community.

RQ3.1 Is it possible to infer user's missing opinions based on his community? If so, what can be considered as the closest proxy?

H4 Some users are more predictable than the others.

RQ4.1 What are the key factors that make a user predictable? Can a "profile" of such predictable user be meaningfully summarized?

RQ4.2 Is it feasible to determine one's predictability beforehand? If so, what are the most informative features?

H5 All proposed models can be used for data initialization in SVD (singular-value decomposition) matrix approximation, which is a state-of-art technique for image restoration in computer vision community.

R5.1 Does low-rank approximation minimize the disparity between original opinion matrix and restored one, and how does it compare to the one with initialized values? If so, which of the proposed models provides the best initialization?

Chapter 3 addresses these research questions.

Social filtering. Clearly, not all of the selected users, and consequently their content, are equally important to an outer prediction task. We propose a three-step filtering technique for a user-based prediction of social trends.

H6 Ordinary tweets of respective groups of individuals are reflective of the moral and economic state of the society, and can serve as a proxy to socio-economic trends and prospective events.

RQ6.1 Can socio-economic trends and prospective events be predicted based on posts of selected group of relevant users?

H7 Content of some individuals is more indicative of a trend of interest.

RQ7.1 Can posts of a subset of *expert individuals* (defined by the context) result in higher predictive power as compared to the unfiltered content?

RQ7.2 If posts of some users exhibit higher correlation with the trend of interest, what are the best strategies for selecting such relevant people?

Answers to these research questions are discussed in Chapter 4.

The rest of the thesis is organized as follows. Chapter 2 presents the details of users' topical attribution and experimental results yielded for a case study of Twitter investment community. A number of models for interpolation of absent opinions as well as their corresponding performance are discussed in Chapter 3. Chapter 4 introduces the concept of *social filtering* and demonstrates the results achieved for a case study of crime trend prediction. And, finally, our contributions and directions of future work are summarized in Chapter 5.

Chapter 2

Topical Experts Detection

2.1 Introduction

Rapid adoption of online social platforms has drastically changed the landscape of what was perceived as traditional media. New communication paradigm empowers literally anyone to broadcast their message to millions. Breaking news witnessed by locals, reaction towards major events and controversial social issues, perception of brands and political leaders, professional advice, even details of daily routines—all that is constantly shared through social networks and microblogs. Acting as an outlet, they provide a unique opportunity to gaining recognition to these highly dedicated users who devote their time to crafting content of extreme value. And while the latter is curated by a small group of elite individuals, the masses of ordinary users rely on them for disseminating interesting information. Latest survey shows that roughly half of participated Facebook and Twitter users regularly consume news on these platforms [50]. Moreover, content producers themselves actively exploit this medium: 54% of US journalists report to find their stories on Twitter, and 79% monitor social media for breaking news [102].

However, absence of content verification on such systems makes them vulnerable to spread of rumors and misinformation. Hence, the task of identifying individuals authoring credible and engaging material is of utmost importance. While opinions of famous users are often in the spotlight, we are interested in discovering individuals that are *experts* in particular fields. Since the notion of knowledge exists only within a specific context (at most couple of them), we aim to detect *topical experts* as opposed to influentials famous across different communities.

While some platforms allow users to explicitly sign up for groups establishing communities based on their interests, others do not support such functionality. For example, to overcome this issue on Twitter, people form implicit networks by following alike [99]: politicians subscribe to politicians, journalists listen to journalists, entrepreneurs track other successful peers, and so on. Automatic detection of such groups has a number of applications including professional follower recommendation, extraction of up-to-date trends in specific domains, finding reliable business intelligence sources, surveillance of suspicious activities, and others.

Although most of the research effort in expertise localization has focused on generating small list of the most influential people with respect to the field of their knowledge, mostly to be consumed as a source for a follower recommendation, our objective is different. Since we are interested in obtaining a representative sample of users whose collective opinions would then be used in an outer prediction task, we would like this pool to be as much diverse in terms of the level of their expertise as possible [87]. That is within the *social filtering* framework detecting individuals with average or even marginal expertise is equally important to identifying most knowledgeable users. In this work, we propose an automatic approach to topical community detection in social media platforms, such as Twitter.

The rest of the chapter is structured as follows. Existing work on detection of

expertise and topicality is reviewed in Section 2.2. We define *profile-* and *behavior-based* models as well as naïve filter in Section 3.3. Experimental setup and results for a case study of investment community identification on Twitter are reported in Section 2.4. And finally, we discuss the implications and directions of future work in Section 3.5.

2.2 Previous Work

Significant research effort has been undertaken with respect to topical experts identification. Very first works were concerned with experts retrieval from knowledge databases which were manually curated within organizations [11, 106]. Since then focus of the research has shifted to mining of existing documents authored by potential candidates, this way allowing to detect qualified individuals without a need in constructing skill databases.

Number of diverse media channels has been explored for this purpose: starting with enterprise corpora [8, 44], followed by discussion groups [96] and Q&A communities [73], with most of research attention concentrated lately on various social media. As opposed to enterprise emails and document repositories, content generated on online social platforms is publicly available, thus making this source extremely attractive for research community. Besides, different types of user interactions supported by these systems provide a possibility to tap into collective opinion of studied community and judge which individuals are perceived as knowledgeable and which are not.

Most of the previous works define experts within the context of one or several areas of their proficiency, thus differentiating between topicality and expertise. While some studies, which treat topical relevance simply as a query matching task, show a

sufficient performance [15, 44, 83, 96], other use more sophisticated approaches. For instance, Balog *et. al.* [8] developed two generative probabilistic models to associate user's expertise with topics directly or through a latent variable represented by a document. Weng *et. al.* [99] and Wagner *et. al.* [93] showed that latent Dirichlet allocation (LDA) is capable of distilling meaningful topics from various Twitter content, such as posts, users' biography and list metadata. Another perspective on user's topical attribution was studied by Lehmann *et. al.* [58]. Authors aimed to classify topically-focused news curators and those with a broad spectrum of interest, with underlying idea that the former can be the best source of authentic and high quality information.

However, most of the research endeavor has focused on modeling user expertise rather than their topicality. Various approaches have been proposed, which can be roughly divided into *network-based* and *feature-based*. The former are variations of PageRank that intend to propagate leaders' influence through explicit or implicit networks. They exercise the intuition that users interacting with authorities would have higher influence than those without such a tie. Indeed, Bhattacharya *et. al.* [12] have shown that domain experts form a strongly connected component, and most of them can be found within 2-hop network of individual authority. LeaderRank proposed for social bookmarking by Lü *et. al.* [66] has shown to outperform PageRank even for noisy data. Weng *et. al.* [99] presented topic-aware modification of PageRank for Twitter. Cheng *et. al.* [22] moved forward and compared how the results vary if expertise is propagated not only through explicit friendship network, but via list-labeling associations. Another interesting version was proposed for knowledge communities [96] by Wang *et. al.* Users were associated with each other through a thread if one of them had voted for another's reply. Authors introduced a weight updating scheme to neglect the impact of colluded groups, and tested number of

strategies for combining relevance and authority. However, such approaches normally favor general authorities and may easily overlook newly joined users. Also they are computationally expensive, thus it might be unfeasible to apply them in the real-time scenarios.

The feature-based techniques have explored many dimensions of candidate representation, covering user-attributed content, structure of his immediate network, engagement with the platform and patterns of his temporal activities. Various aspects of content were considered in the light of user's knowledge, apart from semantics: authorship of the content associated with user (candidate-generated or produced by immediate network) [15], its type (tweets, biography, lists subscribing to candidate user) [93], orientation (original, conversational, reproduced) [58, 74, 78], and even user readability [78] and self-similarity [74]. As for a social graph, feature engineering revolved around the concept of user's leadership, mainly represented by his visibility, mention and retweet impact, information diffusion and other conventional metrics [58, 74, 78]. Some studies considered both static and dynamic social networks [12, 78]. For example, Bhattacharya *et. al.* [12] discovered that domain experts prefer passive following of each other's updates to direct interaction. Additionally, the way users engage with a platform was shown to be expressive of their knowledge and interest in specific topic [15, 58, 78]. Pal *et. al.* [73] have demonstrated that users' expertise in Q&A communities can be successfully predicted based on a number and frequency of answers submitted during early weeks on a platform.

Other works studied how quality of expert identification varies over different media. Guy *et. al.* [44] provided a comprehensive overview of enterprise social applications, such as blogs, wikis, bookmarking, file sharing and others, and showed that profile tags and microblogs demonstrate superior performance. It is expected, since descriptive tagging within an organization would normally boil down to major set of

skills that colleagues believe candidate to posses. However, another cross-platform study on potential of LinkedIn, Facebook and Twitter in detecting knowledgeable individuals [15] showed quite surprising results. Not only LinkedIn was outperformed by Twitter, it showed the worst results among the channels. Moreover, Twitter has proven to be more informative than all three platforms combined. We believe that the cause of it lies in a short nature of Twitter updates (not more than 140 characters), which is expressive of dynamics and persistence of topics in candidate's conversations. It also supports user connections based on their affinity, yet without requiring these ties to be reciprocal. Based on discussed considerations, in this work we concentrate on the Twitter as our target platform.

A different line of research has actively explored this medium as well, particularly Twitter mechanism of list subscriptions [12, 22, 42, 93]. The latter allows users to group accounts they follow into some meaningful categories and provide them with descriptive annotations (e.g. finance, social computing, python development, *etc.*). The assumption is that community itself will discover most prominent individuals, if sufficient amount of users list them under the same or similar areas of expertise. In fact, Wagner *et. al.* [93] have revealed that Twitter lists are more informative about user topical knowledge than his tweets or short biography. Partially it can be attributed to the fact that users not always specify information directly related to their occupation in their bio, or sometimes provide a misleading description as a joke [42]. One more dimension of expertise—locality—was studied by Cheng *et. al.* [22]. Authors also used Twitter lists as the source, but they argued that ability to attribute experts to specific location would provide more benefits for a recommendation (e.g. recommending best lawyers in Houston would allow users seeking for professional advice connect with these specialists directly; recommending most renowned local gourmets can provide an opportunity to discover new restaurants, *etc.*). While work-

ing with Twitter lists yields significant results even for niche topics [12], such strategy also has its limitations. Namely, list-based approaches will fail if they need to discover experts that recently joined the network or motivated but novice individuals. Note that within *social filtering* framework we do not want to bias our sample to the most distinguished practitioners, to the contrary, we would like this pool to express as much diverse opinions as possible.

Finally, we need to point out that there is a number of proprietary services for influence discovery, such as Klout[1], PeerIndex[2], Kred[3], Wefollow[4] and Twitter's own Who to Follow[5]. However, the problem with them is that, first, most require user to explicitly sign up to be discovered by their algorithm, even though the authority is gauged based on the activities on Twitter, Facebook, LinkedIn and other social networks, and second, the details of underlying implementation are not revealed to broad public.

Clearly, judging about individual's expertise is non-trivial, and actually very subjective task. For that reason most of the studies we discussed required vast amount of human participation: either peer-reviews were conducted to evaluate quality of expert selection, or subset of sampled specialists was asked to perform self-assessment. Interestingly, some of the works [44, 58] substituted notion of expertise with interest in topic, making the point that the latter is more feasible to assess, and it is a sufficient proxy to one's knowledge. In this work, on the contrary, we are more interested in defining user's topical attribution rather than exact level of his expertise. As opposed to the works simplifying detection of user's professional domain to query

[1] `klout.com/corp/about` (2015-05-06)
[2] `peerindex.net/about.php` (2015-05-06)
[3] `kred.com` (2015-05-06)
[4] `wefollow.com/about` (2015-05-06)
[5] `support.twitter.com/articles/227220-about-twitter-s-suggestions-for-who-to-follow` (2015-05-06)

matching, we propose an automatic approach to topical community detection based on semantics prevailing within a group of interest. We present a detailed description of aforementioned method in the next section.

2.3 Our Approach

The community detection task is gaining more and more interest due to the recent growth of social media usage in a variety of professional areas. Networks that were initially aimed for mundane communications, now are flooded with business transactions, targeted advertisement, dating opportunities, discussion boards for dedicated topics, and even illegal activities. One notable example is Twitter: its limitation on the post size forces users to produce more concise and informative content, this way making it a perfect medium for getting instant updates from various social circles user is involved in, including professional. However, discovery of such groups in systems which do not have an explicit mechanism of community membership can be a challenging task.

In this work we focus on identification of users' topical attribution based on the content they authored, and select Twitter as our target platform. We cast this problem as a binary classification of users with respect to domain of interest: a user can be relevant or irrelevant to selected domain. We hypothesize that each of these communities has a unique semantic signature: experts normally use shared lexicon, limited number of topics and even the same style of writing. Yet detection of such groups is not easy to generalize, since this task normally requires to have a background knowledge about the field.

Here, we present two models completely decoupled from the area of interest. They only require a positive example of the language used within a community, while

the negative samples are generated automatically. Moreover, we do not impose any limitations on the medium positive samples originate from—it could be anything starting with news articles, blog posts, technical reports, interviews or even chapters from a textbook. We exercise the following intuition for obtaining negative training set: deciding whether a user belongs to specific community in a huge social network like Twitter is a typical example of classification in extremely imbalanced setting. With positive class being underrepresented, a probability of a randomly picked user to be considered as irrelevant is very high. Based on this premise Twitter streaming can be used to populate a negative set. Although there is an infinitesimal chance of a user being mislabeled, we believe our models to be robust to a small percentage of such errors in the training set. Henceforth we refer to the negative dataset as *white noise* and to positive—as *golden* set.

We devise two models—*profile-based* and *behavior-based*—to capture different aspects of user's engagement with the topic of selected domain. Former aggregates all tweets posted by a user in a single *profile* which is considered to be a representative proxy to user's interests. However, it is not able to differentiate between individuals who dedicate significant amount of their content to the topic of studied community and those rarely posting tweets relevant to the group. Also with *profile-based* model it is not clear how the decision should be made when considering a user who joined the group recently or the one who seems to have lost interest in studied topic (a user who stopped posting relevant content). To address these issues we develop a *behavior-based* model which defines topical relevance on a tweet level. We also propose a baseline dependent on the field chosen for a case study. We briefly discuss the domain selected for our experiments and then provide a detailed description for each of the proposed models.

Selected domain. Inspired by recent works that leverage experts' opinions for

Table 2.1: Part of a sample timeline of a stock market expert from *target* Twitter dataset

From the daily chart, 625 in $AAPL wouldn't completely destroy the LT uptrend
In a few $SPY 145.5 weekly puts for 86c
Rainy mornings always throw my clock off for a bit
I'm up 103% and taking some cushion on the $SPY 145.5 puts I have
Drivin' along in my automobile.... http://t.co/HphiN1VS
A zoo. Kids request
So I guess the unlocked $YELP shares aren't immediately for sale

stock market prediction [9, 46, 60, 95], we concentrate our attention on the investment community of Twitter. Obviously, most of the conversations revolve around performance of specific stocks, denoted by *cashtags*—ticker symbols preceded by a dollar sign (*e.g.* $AAPL, $MSFT, $TSLA, *etc.*). Oftentimes practitioners summarize their speculations with trading recommendations, such as BUY, HOLD or SELL. There is also a convention to finish stock-related tweets with a double dollar sign ($$), which is followed by many users. Sample timeline of a target expert can be seen in Table 2.1. Note that since Twitter does not restrict users to sharing professional content only, those active practitioners would also have a significant number of posts related to personal matters. Actually, for half of the stock market experts in our Twitter dataset the fraction of professional content would not exceed the level of 0.33 (see Figure 2.2b). This potentially may lead to a confusion in our models, but we discuss this implication in more details in Section 2.4.

We should discuss the choice of positive and target datasets as defined by our framework. As stated earlier, our goal is to identify topical groups on Twitter, thus *target* dataset is comprised of timelines of stock market experts active on Twitter. We describe how this dataset is obtained in the next section. Negative set, or *white noise*, is populated by selecting random users from Twitter stream. As can be seen from a sample timeline of such user (presented in Table 2.2), their posts cover topics

Table 2.2: Part of a sample timeline from *White noise* dataset

iPad 3 replacement screen from Apple was $299 ! So i just found a place on Yelp that does repairs and got it done for $80 !
"We're the alchemists of our age, we don't know what we need to know yet" - @jjacoby #BICDCon
The dancing ghost loading animation in the Snapchat app is so amusing! http://t.co/x7WHjpsw7c
Martini's with Lisa (@ Holiday Inn Mart Plaza - Cityscape Bar) http://t.co/dZ84lOuwZc
Appreciate the little things in life
If you only knew...

of general interest.

For the *golden*, or positive set, although there is a variety of sources to choose from, such as stock market discussion boards, individual blogs of selected advisors, news articles from financial data vendors (*e.g.* Bloomberg or Thomson Reuters, *etc.*) and many others, we restrict ourselves to *Stocktwits*[6]—microblogging service for investment community. The platform pretty much resembles Twitter with the main difference that users are solicited to share their insights on financial matters. Albeit there is no penalty or moderation of the content which is (slightly) off the topic (e.g. sometimes users wish each other a good weekend or productive week), experts prefer to use the medium for professional conversations only. Thus we believe that negligible fraction of irrelevant posts would be automatically discarded by the volume of financial content. Majority of Stocktwits users have at least 80% of their timelines considered to be relevant to the community (see Figure 2.2a). Similarly to Twitter, users of Stocktwits are limited to 140 characters to express their thoughts. Citation of a ticker by a cashtag is also strictly followed on this platform. For sample posts and user interface see Figure 2.1.

[6] `stocktwits.com/about` (2015-05-18)

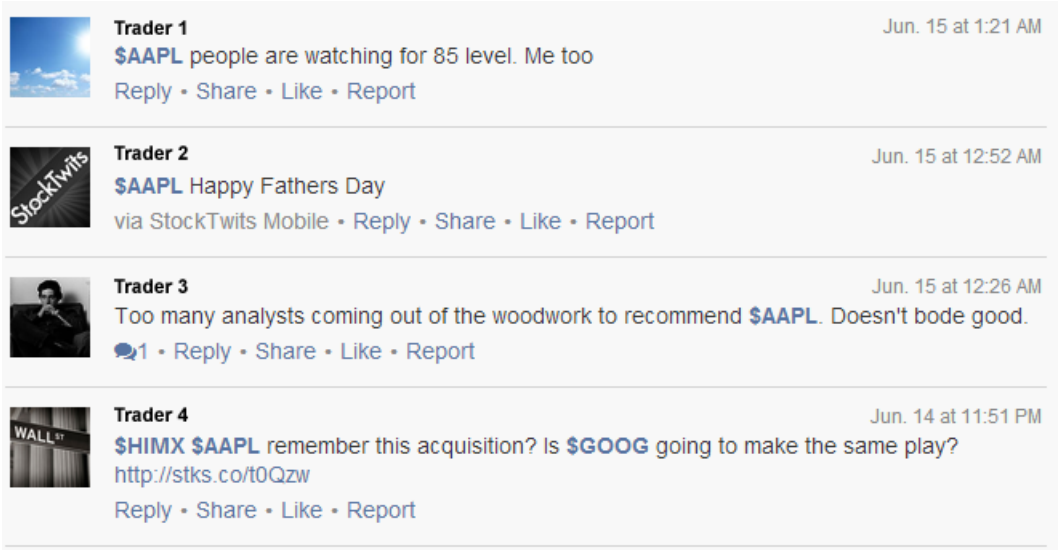

Figure 2.1: Example of Stocktwits posts. Real usernames and avatars are replaced.

2.3.1 Profile-Based

Let $T_i = \{t_1^{\langle i \rangle}, t_2^{\langle i \rangle}, \ldots, t_{N_i}^{\langle i \rangle}\}, N_i \in [1, N]$ denote a timeline of a user i, where $t_j^{\langle i \rangle} = \{w_1^{\langle i \rangle}, w_2^{\langle i \rangle}, \ldots, w_{M_{ij}}^{\langle i \rangle}\}, M_{ij} \in [1, M]$ is a tweet consisting of unigrams $w_k^{\langle i \rangle}$. Then *user profile* U_i is represented by a binary vector space model $U_i = \left(w_1^{\langle i \rangle}, w_2^{\langle i \rangle}, \ldots, w_{|V|}^{\langle i \rangle}\right)$, where $V = \bigcup_{i,j,p} w_{jp}^{\langle i \rangle}, p \in [1, M_{ij}]$ is a global vocabulary and

$$w_k^{\langle i \rangle} = \begin{cases} 1, \text{ if } \exists\ j \colon w_k^{\langle i \rangle} \in t_j^{\langle i \rangle}, \\ 0, \text{ otherwise} \end{cases} \tag{2.1}$$

Each user U_i is associated with a label $L_i \in \{0, 1\}$ with $L_i = 1$ indicating a user relevant to community. The *profile-based* model then uses samples U and corresponding labels L to fit a classifier.

In this model user is represented by a static *profile* compiled by aggregating all his tweets into one document. We posit that such approach allows to detect salient topics[7]

[7]Here and after in this section by the topic we mean topical (domain) orientation of the community to be detected.

across user's lifetime on the platform. We restrict ourselves to unigrams instead of using more sophisticated language model, because, first, all three datasets—*golden*, *white noise* and *target*—exhibit different sets of topics (for *golden* they would be centered around performance of different types of equities, for *white noise* it would be a wide set of general topics, while the *target* would combine both), thus it is not valid to extract topic-words distributions jointly from these sets, and it might lead to a meaningless result; second, since all professional groups are known to use jargons, which usually consist of one word, the presence (or absence) of such lexicon could discriminate between relevant and irrelevant users. However, this model can be easily extended to higher order representations.

2.3.2 Behavior-Based

In the *behavior-based* model user U_i is represented as a collection of his individual tweets $U_i = \{w_{km}^{\langle i \rangle}\}, k \in [1, N_i], m \in [1, |V|]$, and $w_{km}^{\langle i \rangle} = 1$, if $w_{km}^{\langle i \rangle}$ has occurred in $t_k^{\langle i \rangle}$. To capture changes in behavior of the user with respect to topic of interest, *behavior-based* model builds the classifier for individual tweets as opposed to full timeline as in *profile-based* model. We also define vector of labels $l_i = (l_1^{\langle i \rangle}, l_2^{\langle i \rangle}, \ldots, l_{N_i}^{\langle i \rangle})$, where $l_k^{\langle i \rangle}$ is associated with each individual tweet k for the given user U_i. For the training $l_k^{\langle i \rangle} = L_i$. To predict whether U_i belongs to community or not, we calculate $r_i = \frac{1}{N_i} \sum\limits_k \hat{l_k^{\langle i \rangle}}$—a ratio of tweets considered as relevant by aforementioned classifier. Then if it exceeds a threshold θ, users is considered relevant:

$$\hat{L_i} = \begin{cases} 1, \text{ if } r_i > \theta, \\ 0, \text{ otherwise} \end{cases} \tag{2.2}$$

We would like to point that *behavior-based* model does not consider actual timing of the posts, which could have potentially shed some light on detecting users who just

showed interest in the area and those who lost it long ago. However, it is able to reflect the notion of commitment—to what extent target user is dedicated to producing content related to the field of interest as opposed to the one covering broad set of topics. We expect this model to outperform *profile-based*, since it can reduce the rate of false positives, which are unavoidable in case of a static user representation. It is clear that it carries a significant computational overhead, since this method requires every single tweet to be classified by machine learning model.

2.3.3 Domain-Specific Filter

To preserve the accuracy of *behavior-based* model while decreasing running time, we introduce a naïve baseline tailored for this specific domain. *Domain-specific filter* employs the same approach as in *behavior-based* with the only difference that instead of using a classifier it scans for occurrences of *cashtags* (e.g. $AAPL, $GOOG), and marks tweet as relevant if at least one was spotted. That is, if C is a regular expression defining universum of cashtags, then

$$\hat{l_k^{\langle i \rangle}} = \begin{cases} 1, \text{ if } \exists\ m\colon w_{km}^{\langle i \rangle} \in C, \\ 0, \text{ otherwise} \end{cases} \tag{2.3}$$

$\hat{L}_i$ is derived based on threshold θ similarly to *behavior-based* model.

The underlying idea is based on the fact that central object of discourse in this community is a specific company (or companies) represented by its ticker. Although presence of cashtags is not required for a tweet to be considered valid from investment standpoint, precise nature of these conversations results in every active expert listing a specific company at least once. Thus the only part left is to determine appropriate value for the threshold θ. Also since the model does not require actual training, it is expected to have the fastest running time among the three.

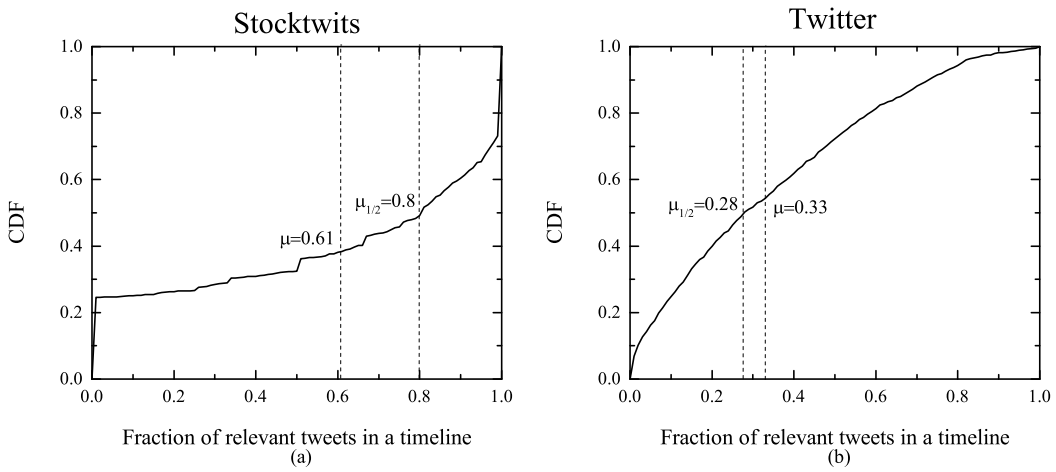

Figure 2.2: Cumulative distribution function of a fraction of relevant tweets in timelines of (a) *golden* Stocktwits and (b) *target* Twitter datasets as determined by *domain-specific filter*. μ and $\mu_{1/2}$ denote mean and median of the distributions.

We have to note that *domain-specific filter* cannot be easily generalized to all possible domains. However, one might propose a slight modification: instead of a regular expression for cashtags one would have to generate a list of top unigrams from a lexicon used in target group and then check for their occurrences.

For this specific domain the goal is to determine whether a naïve baseline powered by a regular expression would suffice for a given task. If not, machine learning approaches introduced earlier should be used.

We also report that results of the baseline conform to our expectations regarding homogeneity of *golden* and *target* datasets (see Figure 2.2), with median of relevant posts equal to 0.8 for Stocktwits and 0.28 for Twitter.

2.4 Experimental Results

In this section we discuss the datasets collection, preprocessing procedures, experimental scenarios and the performance yielded by proposed models.

2.4.1 Dataset Description

Target Twitter dataset [60] was obtained in four steps. First, tweets mentioning 60 tickers from a manually generated list of company names and commonly used synonyms (*e.g.* "Apple Inc", "AAPL", "#AAPL" or "AAPL") were collected using Twitter Search API. Then during March 27 till June 20, 2012, all posts of users that authored first set of tweets returned by Search API were streamed. After that tweets of each individual user were automatically tagged as trading-related if: (i) tweet ended with double dollar sign ("$$"), or (ii) tweet contained at least one ticker and at least one of the predefined action words.[8] Users were considered to be traders only if they pass a threshold based on monthly, weekly and daily frequency of trading-related tweets. We ended up with 6512 users, out of which we randomly selected 1000 to constitute stock market experts dataset.

Golden Stocktwits dataset We randomly chose 1000 contributors and then crawled their timelines since the launch of Stocktwits (May 27, 2008) till March 31, 2014.

White noise Twitter dataset In order to collect representative timelines of randomly streamed Twitter users, we also employed two-step approach. First, we selected users authoring tweets collected from Twitter Streaming API, restricting ourselves to English tweets only. And then we collected historical data of every seed user. We limited *white noise* dataset to timelines of randomly cho-

[8] Action words are defined as verbs occurring in a proximity to a ticker symbol with a high frequency across the whole dataset.

Table 2.3: Statistics on the size of a user timeline for all used datasets

	Stocktwits	White Noise	Twitter
min	1	1	3
median	6	69	3954
avg	173	264	5528
std	812	413	5857
max	13890	2800	48539

sen 1000 users. Note that approach involving Twitter REST API (querying historical timelines) can yield up no more than 3200 of recent tweets per user. Statistics on the size of a timeline for each dataset are presented in Table 2.3. As can be seen, with an average of more than 5K tweets per expert, *target* dataset is represented very well.

2.4.2 Preprocessing

We performed a standard preprocessing procedures on the main datasets: we applied lower case, tokenized the documents into unigrams, removed stop-words, punctuation and digits. We also opted to ignore mentions (e.g. @twitter, @POTUS, *etc.*), hashtags (e.g. #RedNoseDay, #21demayo, *etc.*) and URLs. However, we decided to preserve all individual cashtags with preceding dollar sign being removed. We also applied bandpass filtering on rarely occurring terms.

2.4.3 Experimental Scenarios

We tried binary (described in *profile-based* model) and inverse document frequency (idf) data representations with both of them yielding similar results. In terms of a classifier, we selected support vector machine (SVM) for both *profile-* and *behavior-based* models. Relevance thesholds were empirically set to $\theta_1 = \theta_2 = 0.3$. We further

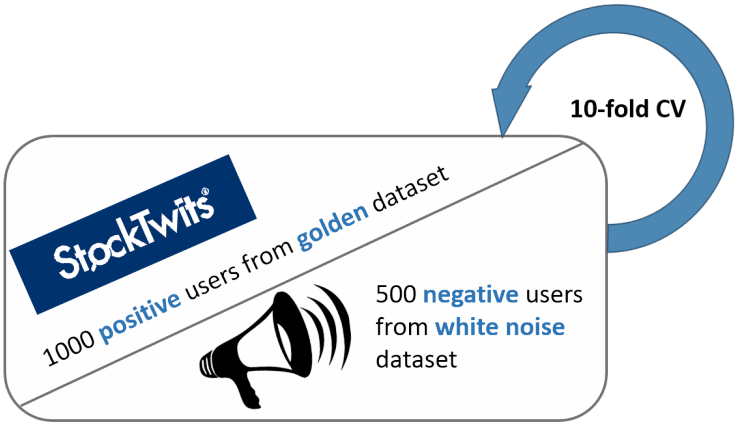

Figure 2.3: Experimental scenario 1.1. Classifier is fitted on *golden* Stocktwits and *white noise* datasets. It is assessed using 10-fold cross-validation.

describe experimental settings used in number of scenarios.

Scenario 1.1 is aimed to evaluate how good the model was fitted. Both supervised models (*profile-* and *behavior-based*) are trained on timelines of 1000 positive users from the *golden* set and 500 negative from the *white noise*. Performance is assessed using 10-fold cross-validation. Basic purpose of such setup is to ascertain that the models are capable of predicting unseen samples of the same type as trained with. Schematic explanation of this scenario is provided on Figure 2.3.

Scenario 1.2 is of utmost interest for us, since it describes the expected *application setting*. Target users are predicted with model learned from automatically generated training set. To make sure that classifiers are not biased towards the positive class, we add 500 negative users from *white noise* dataset to our target test set. Schematic representation of this scenario is depicted on Figure 2.4.

Scenario 2.1 explores the validity of assumption that timelines of target Twitter users are noisy. This setup is similar to scenario 1.1 with the difference that

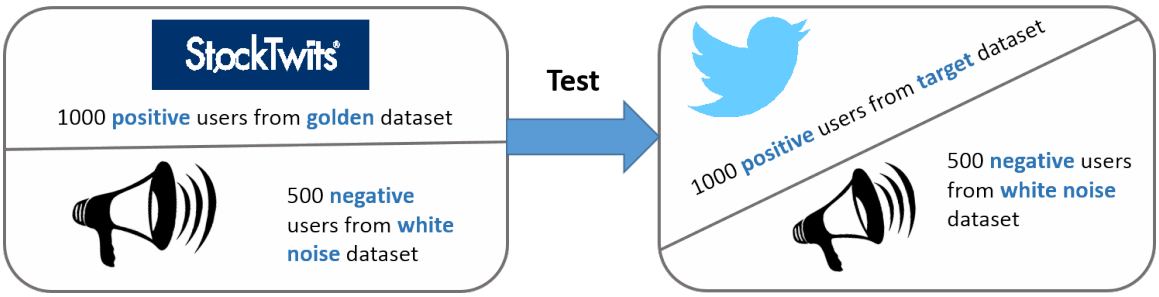

Figure 2.4: Experimental scenario 1.2. Classifier is fitted on *golden* Stocktwits and *white noise* datasets. It is tested on *target* Twitter and *white noise* dataset. It is a target scenario in which model learns from automatically selected data to determine relevance of target users.

golden set is now replaced with *target Twitter* dataset. Goodness of fit is tested using 10-fold cross-validation. We believe that *behavior-based* model will fail, because of confusion introduced by training tweets mistakenly tagged as positive. Recall that the fraction of relevant tweets is at most 0.33 for at least half of studied experts (see Figure 2.2b). *Profile-based* model is believed not to degrade significantly. We elaborate discussion on results obtained for this scenario in the next subsection.

Scenario 2.2 further explores the hypothesis of *noisy timelines* of target Twitter users. As we discussed before, majority of posts in Twitter timelines of selected experts is concerned with matters irrelevant to studied community. We believe that it might negatively affect the performance of application scenario. To validate the hypothesis, in this experiment we consider a setting exactly the opposite to the one described in scenario 1.2. That is *target* Twitter dataset is considered to be sufficient positive sample, and the models learned from it are tested on *golden* Stocktwits set. To make sure that none of the classes is discriminated, both training and test sets include negative users. We expect both machine learning models to fail in this experiment with *behavior-based*

showing significantly worse performance.

Domain-specific filter is employed on the same testing scenarios with the difference that the actual training phase is omitted.

2.4.4 Discussion

Performance across designed scenarios

We report the results obtained for 4 validation scenarios in Table 2.4. We list average, positive and negative F-measure, as well as accuracy, precision and recall.

Expectedly, all models including *domain-specific filter* performed well in cross-validation scenario when trained on clean data (scenario 1.1). Skewed class distribution (with number of negative samples equal roughly to half of positive) did not have any negative impact on either of experiments. *Behavior-based* model showed even better performance ($F_1 = 0.96$ compared to 0.94) when tested on the target set (application scenario 1.2). Both *profile-based* model and *naïve filter* degraded significantly, yet achieving decent results ($F_1 = 0.78$ and $F_1 = 0.66$ respectively).

Interesting insights can be derived from results obtained in scenarios 2.1 and 2.2. As we expected, *behavior-based* model failed to cope with mislabeled training data: $F_1^- = 0.16$ and $recall = 0.54$ support the assumption on prevalence of false positives. Cross-validation scenario on *target* and *white noise* datasets did not affect performance of *profile-based* model, however, replacing positive test set by Stocktwits led to same consequence as for *behavior-based*.

We would also like to point out that in all scenarios where training data was not corrupted machine learning approaches have beaten the baseline.

Table 2.4: Comparison of the results yielded by proposed models. F_1^+ and F_1^- stand for F-measure of positive and negative classes respectively. PB denotes *profile-based* model, BB—*behavior-based* model and DSF—*domain-specific filter*. In scenarios 1.x model is trained on golden set and *white noise*, in scenarios 2.x—on target set and *white noise*. Scenarios x.2 denote cross-validation, and x.1—testing on the target and golden set correspondingly.

	Model	Scenario 1.1	Scenario 1.2	Scenario 2.1	Scenario 2.2
F_1	PB	0.94	0.78	0.99	**0.52**
	BB	0.94	**0.96**	**0.49**	**0.48**
	DSF	0.81	0.66	0.70	0.81
F_1^+	PB	0.96	0.81	0.99	0.45
	BB	0.96	0.97	0.81	0.81
	DSF	0.84	0.65	0.71	0.83
F_1^-	PB	0.91	0.74	0.98	0.58
	BB	0.93	0.95	**0.16**	**0.14**
	DSF	0.78	0.66	0.69	0.78
Accuracy	PB	0.94	0.78	0.99	0.53
	BB	0.95	0.97	0.70	0.69
	DSF	0.81	0.66	0.70	0.81
Precision	PB	0.95	0.78	0.98	0.70
	BB	0.93	0.95	0.82	0.82
	DSF	0.82	0.75	0.76	0.82
Recall	PB	0.93	0.82	0.99	0.64
	BB	0.96	0.91	**0.54**	**0.54**
	DSF	0.72	0.74	0.77	0.86

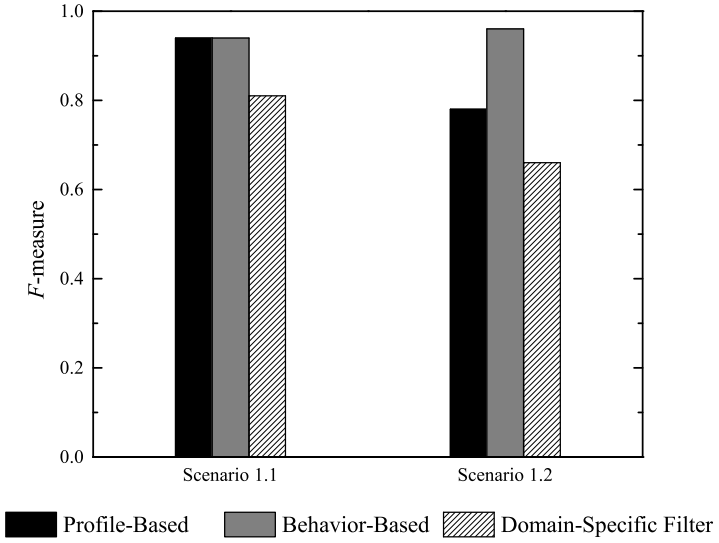

Figure 2.5: Comparison of the results yielded by proposed models. Models are fitted on *golden* Stocktwits and *white noise* datasets. In scenario 1.1 they are tested using 10-fold cross-validation, in scenario 1.2—*target* Twitter and *white noise* datasets.

Performance in application scenario

We now discuss how performance varies across different models for our aimed scenario 1.2 (see Figure 2.5). Although the *behavior-based* classifier has significantly outperformed other models due to its capacity to capture dynamics of topic usage by an expert ($F_1 = 0.96$), we cannot apply it in a real-time setting because of extreme computational overhead. *Naïve baseline* can be used for tasks requiring testing of a vast amount of users—for a case of cashtag-based regular expression it achieves relatively decent performance of $F_1 = 0.66$ and the best execution time among the models.

For trivial scenarios which require higher accuracy, *profile-based* model seems to be the most suitable candidate: both time-wise and performance-wise it resembles a trade-off between first two.

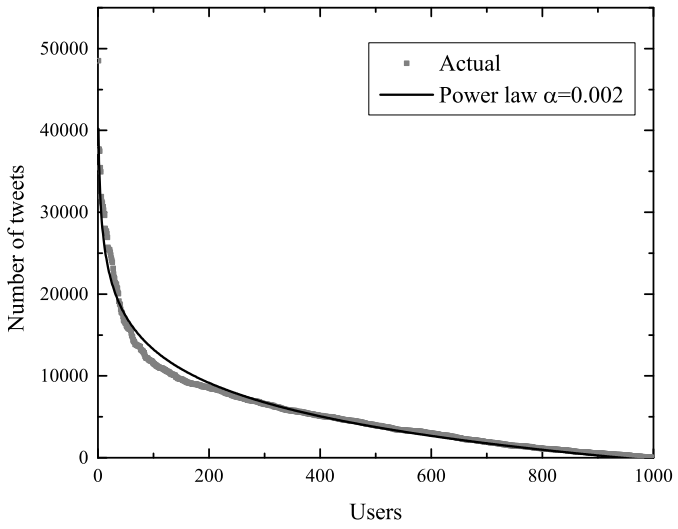

Figure 2.6: Distribution of content generation in *target* Twitter dataset

Dependency on the availability of posts

One legitimate question stems from the fact that users are considered within the context of their timelines. It is interesting to know how availability of their posts can affect the quality of prediction. Specially, considering the fact that content generation in the *target* dataset is described by power law (see Figure 2.6), even though all users are coming from the same homogeneous community, and data collection was not discriminating less active users.

Here we speculate how dependent devised models are on the timelines with a bigger size, or are they at all. We plot the prediction accuracy of target expert users (negative users are ignored in this setting) on Figure 2.7. Surprisingly, none of our models seem to rely on the size of user timeline available for testing. That is all models predict the relevance of users with around only 40 available tweets with the same accuracy as those with more than 5K tweets. It leads to a significant implication

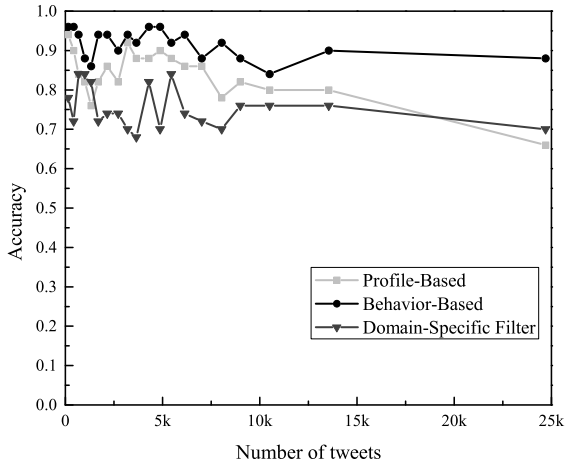

Figure 2.7: Dependency of accuracy on the size of available timelines when predicting target users from Twitter dataset

indicating that all of these models can be successfully used even on candidates with extremely small portion of a timeline observed, which means that even for those users who produce tremendous amount of information, we only have to analyze its small fraction to make a correct judgment.

2.5 Conclusion

Summary

In this work we proposed an automatic approach to discovery of expert's topical attribution in social networks which do not allow its users to form explicit groups based on their interests. Presented approach exploits user authored content as a proxy to his interest. We casted this problem as a binary text classification task, and exercised the intuition that people within the same community would share the same semantic patterns. We described the procedure for automatic acquisition of training data based on the concept of extremely imbalanced binary classification. Our models

require only a positive sample of a language used in the domain of interest with no restriction on the source this data is coming from. Negative examples are created automatically by randomly streaming Twitter feed. This way, unlike in other works, our framework does not need human participation (neither for annotating of the training set or for evaluating results).

We devised two machine learning models—*profile-based* and *behavior-based*—which capture respectively static and dynamic components of user engagement with the topic of interest. We also proposed a *domain-specific filter*—a baseline tailored to specific domain used in our case study. With a slight modification it can actually be extended to other domains.

Experimental results for investment community of Twitter have shown that all three models yield decent performance for a targeted application setting: with *naïve baseline* running in linear time and achieving $F_1 = 0.66$, *behavior-based* obtaining the best results ($F_1 = 0.96$) but being computationally expensive, and the *profile-based* being a trade-off between these two both from time and performance points of view.

We also have discovered that none of these models relies on the size of the timeline of a candidate user, meaning that only fraction of posts can be analyzed even for those very active individuals, this way saving time yet providing declared level of quality.

Implications

This framework can be successfully applied to automatic discovery of topical groups on platforms, such as Twitter. Set of selected candidates can be then used for a tailored professional recommendation or selection of an "expert crowd" relevant to external analytical task. For example, for the reported case study, content of such experts can be simply treated as a set of recommendations which can be used for devising a sophisticated trading strategy.

Future work

Many avenues of research can be considered in the future work. For instance, higher-order models can be explored in order to represent deeper semantic attribution, alternatively variation of traditional topic models can be employed to learn disjoint set of topics characterizing the community. Also *behavior-based* model can be modified the way it actually incorporates explicit temporal analysis. Another way for improvement could be designing of a model hybrid of *profile-based* and *behavior-based* to improve performance of the former and running time of the latter. Clearly, all proposed models have to be validated on a different target community. And finally followup research can focus on discovery of users' actual expertise.

Chapter 3

Interpolation of Missing Opinions

3.1 Introduction

Identification of credible stories and their sources is a task of paramount importance not only for professional news discovery[1], but virtually for any research aiming to derive conclusions from user-generated content. Models based on corrupted and noisy data can significantly degrade in performance or even deliver plausible but erroneous insights. However, approaches with user-centric analytics in their core can easily minimize harm and confusion caused by misleading or spammy user by explicitly damping his impact or even by fully discarding his content. For instance, models predicting stock market movements based on recommendations of selected group of experts have drastically outperformed their counterparts using bulked data [9, 46, 60, 95]. Intuitively, while some applications can offer satisfactory results examining aggregated content, others require a proxy to user's credibility, authenticity and intentions. We focus on the latter group of approaches.

The problem with them, though, is that when a group of "expert" users is iden-

[1] `storyful.com/about/` (2015-06-01)

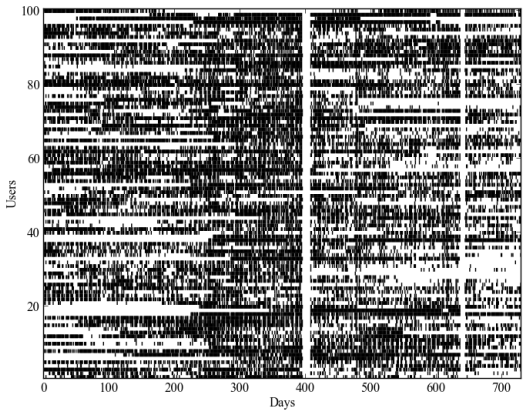

Figure 3.1: Rastergram of daily opinions shared by 100 most active users of Stocktwits Apple community captured during 730 days of observation. Black vertical bar indicates that a user had at least one post on that day. Ideal data would resemble a black square.

tified, a new challenge arises. Techniques directly working with user streams are inherently prone to vast amounts of missing data. No matter how productive selected users are, activity gaps in their timelines are inevitable (see Figure 3.1). Here, by expert users we mean individuals as opposed to celebrities, news agencies or big companies whose corporate accounts in social media are normally managed by a group of employees. Periods with absent data can be attributed to a number of reasons, starting with sampling occurring during data collection, users simply not having access to the medium or even intentional self-censorship [32]. Traditional solutions to this problem either arbitrarily fill out missing data or completely ignore such periods. Both carry their own implications—first one introduces additional noise which can result in error propagated to the main model, and second does not take advantage of content generated by others.

But what if we can infer what a user was up to when he was offline, or if he deliberately opted out to share his insights? Our objective here is to assess the fea-

sibility of interpolating so-called *missing opinions* for individual users based on their previous activities and meta-information extracted from the network. We exercise the intuition that when publishing content, especially opionated, users are influenced by external events (news), ideas and judgments of their friends and, of course, their views, biases and interests. Motivated by aforementioned studies predicting market performance based on collective opinions, we similarly focus on the Twitter-like financial community. While we are not interested in forecasting market behavior, or even more—designing profitable trading strategies executing ideas about market performance—we find this community to be a great fit for a given task. First, this group of individuals is extremely homogeneous—they share common goals, motives and interests; also publisher-subscriber relationships on this platform carry rather practical meaning (a user would typically follow most renowned and successful peers) than represent organic friendship like other platforms do[2]. Second, even though some of the posts have simply reporting character, most of them are posed as recommendations which can be read as one's predictions for the near- (or long-) time market performance.

Our contributions are three-fold: (i) we propose a supervised model for classifying user recommendations with respect to prospective market movements as *bullish* (prices are expected to rise) or *bearish* (conversely, investors expect to carry losses); (ii) we devise *opinion-based* and *content-based* models for interpolation of missing opinions as well as a number of baselines, and show that while being sufficient on their own, when applied in a state-of-art technique for initialization they significantly improve the quality of inference; (iii) we analyze how quality of behavior prediction varies across different groups of users and examine the features informative of highly

[2]However, we do not deny neither ties between practitioners who personally know each other, nor a friendship developed as a result of on-platform communications.

predictable users.

The rest of the chapter is organized as follows. We review the previous work in the field of user behavior modeling with respect to content generation in Section 3.2. We discuss problem statement, specifics of selected domain and details of proposed models in Section 3.3. Dataset description, details of evaluation and discussion of results are presented in Section 3.4. Contributions, implications and future directions of research are outlined in Section 3.5.

3.2 Previous Work

A great body of work has been devoted to modeling of user behavior in various contexts. Here, we focus only on the research touching user modeling with respect to the content he generates and interacts with.

Significant attention has been paid to the application of content recommendation. Collaborative filtering techniques, offering outstanding performance when recommending canonical types of items (*e.g.* products, movies, books, *etc.*), failed when were applied to a content with extremely short lifetime. For instance, news, which usually are relevant for one day at most, simply do not have enough time to collect amount of ratings sufficient for a decent personalization. This resulted in emergence of content-based approaches that derive patterns of engaging content from user's intrinsic interests, trending topics and other proxies. Liu *et. al.* [62] proposed a content-based model as an extension to existing collaborative filtering approach for recommendations on Google News. Authors extracted genuine interests of individual users from click logs and combined them with regional news trends, which significantly increased articles click-through-rate and overall visiting frequency of the portal. They have also shown that user's interests tend to change over time, which has to be ad-

dressed when designing such models. Another works [3, 107] selected Twitter as a source of upcoming news. Yin *et. al.* [107] proposed a mixture model similarly capturing both personal interest and timely content. Also, in order to alleviate high impact of the latter component, authors proposed a penalization of popular items based on inverse entropy. Abel *et. al.* [3] narrowed down the problem of interest discovery to automatic generation of user profiles based on the content they published. Authors have shown that proper choice of profile representation (entities, topics, hashtags) as well as considered timespan (weekend *vs.* weekday, recent *vs.* whole) has a significant impact on recommendation results.

Another factors affecting user's reaction to content—its visibility and exposure time—were considered with respect to search behavior [110] and story promotion on Digg [48]. Zhang *et. al.* [110] modeled click behavior in a context of a user performing specific task during subsequent query sessions. User was treated as an agent transitioning between different states based on the relevance of provided search results and probability of seeing a document earlier. Authors have found that in such setting users tend to click more during subsequent queries as they refine its formulation, and also ignore the documents shown in results before. Hogg and Lerman [48] used analogous framework to model user voting patterns on Digg. They have revealed that apart from user's exposure to a story it is crucial to differentiate between users based on their personal relationship with story submitters and supporters. Authors showed that using such approach popularity of a story can be successfully predicted only based on the early votes.

Number of works studied how continued participation can be predicted based on user's previous behavior and interaction with others. DeDeo [34] considered cooperative edits of Wikipedia articles both in the light of individual contributions and collective effort. Wilkinson [101] analyzed the patterns leading to the final point of

user's activity on different peer production platforms. Joyce and Kraut [52] studied how different reactions towards the first post by a newcomer of online support group can affect his subsequent activities. Interestingly, authors have found that not only the response of community members, but also characteristics of the initial post itself are informative of user's decision to contribute again. Althoff and Leskovec [5] explored a problem of donor retention in a crowdfunding community from perspective of volunteer's historical activity, quality of communications with an individual requesting the donation, and also attributes of specific project that volunteer donated to, such as its success and cost.

Information flow in the network and effect of influence on social friends is another facet considered in user modeling. Ver Steeg and Galstyan [85] proposed a transfer entropy model to show causal relationship between timings of tweets containing URLs posted by pairs of users. More sophisticated approaches have also examined the content and user's propensity to share the posts of their friends. For instance, State and Adamic [84] studied spread of support for same-sex marriages on Facebook as a function of user's demographics, personality and level of adoption by their friends. Hogg *et. al.* [49] predicted whether a Twitter user would respond to friend's post discussing controversial topics. They accounted for user's activity, interest in the topic, overall disposition to retweeting content of others, and also proposed an approximation technique to estimate content visibility for different types of users. Xu *et. al.* [103] went even further and presented a mixture topic model for actual generation of new content, not a reply or a retweet. They examined breaking news, user's intrinsic interests and posts of friends as sources motivating a person to publish a tweet. In turn, Artzi *et. al.* [7] studied what characteristics of a user and a tweet itself would make it more likely to receive a response.

While approaches incorporating a knowledge about underlying motives which

drive user's actions show outstanding performance, straightforward techniques can also suffice for particular tasks. For instance, whether a user would submit a post in specific time frame or not can be successfully predicted by observing his temporal patterns and modeling them with point process [31]. Similar technique can be also employed for reconstruction of partially observed temporal networks. Stomakhin *et. al.* [86] applied self-exciting point process for identification of gangs acting as offenders in retaliatory incidents. Authors exercised the intuition that interaction between pairs of rivalry gangs would temporally cluster together. Cho *et. al.* [25] incorporated a spatial component into the previous model and have shown the improvement and also its applicability to prediction of participants and timings of future events. The latter studies are conceptually closer to a proliferating field of link prediction, which is not in the focus of this work. We refer an interested reader to a seminal paper by Liben-Nowell and Kleinberg [61] and surveys on link mining [41, 45].

Another interesting study was presented by Li and Cardie [59]. Authors employed multi-level Dirichlet process for automatic extraction of personal important events from Twitter timelines. But the closest work to ours is by Lakkaraju *et. al.* [55]. They modeled the process of human evaluating various items as a function of his expertise and item characteristics leading to a confusion. Similarly to us they inferred user's interpretation rather than the plain presence or absence of his activity, as it was discussed in other works. However, in contrast with this paper, we are not interested in true label of user's evaluations. Essence of opinion is to carry a subjectivity of individual expressing it, thus the task of assessing its "quality" is not trivial by itself. Although we understand that to some extent opinions can be evaluated using the factual information if there is any reference to external events, judging about the "correctness" of these opinions is not of our primary interest.

As we can see, the existing body of work supports our hypothesis that when mak-

ing a decision to publish some content, user is a subject to an influence by breaking news, opinions of his friends, interplay of both and even their visibility, and, of course, his personal views and interests. Therefore we believe that user's immediate friends and peers from the same community would produce a content which can be treated as a proxy to current events which trigger formulation of his opinions, even when not publicly shared. We discuss details of proposed models in the next section.

3.3 Our Approach

3.3.1 Selected domain

We start by discussing peculiarities of selected domain. *Stocktwits*[3] is a Twitter-like microblogging service for market practitioners. Similarly to Twitter, it imposes 140 character limitation on post length, supports publisher-subscriber mechanism, allows for reposts, replies and likes. Platform solicits its users to publish their opinions and insights with regards to stock market behavior, and although posts on broad topics are usually not moderated, most of the content belongs to financial domain. When discussing specific stocks or particular equities, users are required to denote them with cashtags—ticker symbols preceded by a dollar sign (*e.g.* $FB, $EURUSD, $GLD, *etc.*). This allows users to create their feeds not only based on the people they follow, but also from tickers they are interested in. The latter would additionally render the content generated by people a user is not actually subscribed to. This carries the following implications: first, a user can get an instant access to content of others without publicly showing that he is following them (it gets especially handy when a user wants to get updates from his direct competitors), and, second, it enables fast and easy discovery of experts who recently joined the community. Finally, practitioner can

[3] `stocktwits.com/about` (2015-06-01)

Figure 3.2: Example of Stocktwits posts with *Bullish* and *Bearish* annotations. Real usernames and avatars are replaced.

annotate his posts with optional *Bullish* (optimistic) or *Bearish* (pessimistic) tags, if he wants to explicitly indicate his opinion regarding specific stocks (see Figure 3.2). Although very useful, labeled posts constitute only 20% of all content in our dataset. We refer to *bullish/bearish* posts as the smallest unit of user's opinion. Since the majority of most active contributors post frequently to reflect intraday changes in stock price, it makes sense to consider aggregated opinions to neglect these price fluctuations. We discuss the details of proposed aggregation later in this section.

We believe that selected community is suitable for our task, since, first, it is extremely homogeneous (*i.e.* prevailing topics are shared between most members of the platform, and they are limited to investment), second, users have a common goal of predicting market movements, and third, they analyze recommendations of their peers in order to shape their own opinions. Therefore we hypothesize that one's opinion can be inferred from his historical data and content published within the community. We propose two models—*sentiment-based* and *content-based*—as well as a number of baselines to interpolate missing opinions. We also analyze individual user predictability and explore whether it has any correlation with behavior he exhibits online. And, finally, we examine the applicability of the state-of-art technique for matrix restoration from computer vision community.

3.3.2 Bull-Bear Classifier

Although the main discourse on the platform revolves around prospective market performance, and majority of posts carry one's opinion regarding that, only a small fraction of content has explicit labels indicating sentiment of a user who published it. Therefore, there is a need in inferring a *bullish/bearish* sentiment for posts without provided annotation.

It is a typical example of sentiment analysis problem which can be approached using a lexicon-based technique, commonly adopted by scholars researching on stock market community. However, such approaches are known to suffer weak performance if applied to a very narrow and specific discipline, since the dictionaries consist of general lexica. This issue can be alleviated by developing and using domain-specific dictionaries. Nonetheless, both require a document of appropriate size to deliver satisfactory results. Most of the aforementioned studies work with hundreds of aggregated tweets as opposed to us aiming to define an attribution on a very fine-grained level. Also similarly to Twitter, Stocktwits is known for its bizzare and shallow language, partially developed by users for a sake of brevity; moreover, investment community actively uses its very unique jargon completely incomprehensible for an outsider (see Table 3.1). This makes us employ a supervised approach capable of extracting underlying patterns automatically and not requiring human labor to update existing dictionaries. Additionally, we examine an applicability of lexicon-based approaches (both general and financial) to this task. To the best of our knowledge, it is a first study to classify short segments of investment recommendations (recall that each post is limited to 140 characters) using general and financial lexica.

We use linguistic inquiry and word count (LIWC) [77] as a general dictionary and Loughran-McDonald financial dictionary [65], which was shown to outperform other general lexica.

Table 3.1: Sample Stocktwits posts

$IWM http://chart.ly/ws6jk6g Intraday chart putting a systematic pullback from the Double Top. Offering nice scalp with TZA.
$MBLY Might be getting ready. Momentum indicators starting to turn higher. Keep an eye on this one.
$RDC http://chart.ly/s9wtkxn Price below & descending TL Neckline. 31.80 not out of the picture here. Lots of wk. to do.
"@CaptainJohn: $GOGO twin breakouts! $IWM $RUT $PNQI fasten your belts #GOGO" @dark_trader this ones a Bronco brother thank you!
$LNKD nice squeeze last 3 days.. congrats longs
$EVR has some room to run
$GDX bulls doing a poor job. Still bearish
$GS http://chart.ly/3b8r94n Setting up to B/O above LT TL res, still in channel up, first tgt still around 180
All in place 2. $NYAD $SPX http://chart.ly/6acusoo

Let $p_i = \{w_1, w_2, \ldots, w_{M_i}\}$ denote a post consisting of M_i unigrams, $i \in [1, N]$, where N is the total number of available posts. Then a post i is assigned a $score_i = \log \frac{1+p_i}{1+n_i}$, where $p_i(n_i)$ is a number of occurrences of terms from positive (negative) dictionary in post i. Its sentiment attribution is then derived as:

$$\widehat{l_i} = \begin{cases} -1, \text{ if } score_i < 0, \\ 1, \text{ if } score_i > 0, \\ random(-1, +1), \text{ otherwise} \end{cases} \tag{3.1}$$

We believe that unigrams would also suffice for the supervised model, since the posts are typically short, abrupt and normally do not make a syntactically proper sentence. Let a binary vector space model $\widetilde{p_i} = \left(w_1, w_2, \ldots, w_{|V|}\right)$, where $V = \bigcup_i w_{M_i}$ is a global vocabulary and

$$w_r = \begin{cases} 1, \text{ if } \exists\ j\colon w_r \in p_j, \\ 0, \text{ otherwise} \end{cases} \tag{3.2}$$

represents a post i in a supervised setting. Posts that are manually annotated with a sentiment (*Bearish* or *Bullish*) by a user who published it are associated with a label $l_i \in \{-1, 1\}$. Our supervised model then uses samples $\widetilde{P} = (\widetilde{p_1}, \widetilde{p_2}, \ldots, \widetilde{p_{N_L}})$, where N_L is a number of labeled posts, and corresponding labels $L = (l_1, l_2, \ldots, l_{N_L})$ to fit a model. We expect it to outperform both lexicon-based approaches.

3.3.3 Community-Based Inference

Let $T_i = \{\left(p_1^{\langle i \rangle}, t_1^{\langle i \rangle}\right), \left(p_2^{\langle i \rangle}, t_2^{\langle i \rangle}\right), \ldots, \left(p_{N_i}^{\langle i \rangle}, t_{N_i}^{\langle i \rangle}\right)\}$ denote a timeline of a user i, where tuple $\left(p_j^{\langle i \rangle}, t_j^{\langle i \rangle}\right)$ represents user i's post j along with its timestamp: $t_1^{\langle i \rangle} < t_2^{\langle i \rangle} < \ldots < t_{N_i}^{\langle i \rangle}$.

Let us introduce the smallest unit of aggregation $\Delta\tau$, which is used to stabilize users' opinions across their posts, with $(\tau_1, \tau_2, \ldots, \tau_D), \tau_d = \tau_{d-1} + \Delta\tau$, denoting edges of a sequence of consecutive time bins. Then $U_i = \left(o_1^{\langle i \rangle}, o_2^{\langle i \rangle}, \ldots, o_D^{\langle i \rangle}\right)$ is a *collection of opinions* of user i, where D is a total number of aggregated time bins across all users in the pool. Individual opinion is defined as follows:

$$o_d^{\langle i \rangle} = \begin{cases} \underset{l \in \{-1,1\}}{\operatorname{argmax}} ||\widehat{l_j^{\langle i \rangle}} = l||_0, \text{ if } ||\widehat{l_j^{\langle i \rangle}}||_0 > 0, \\ 0, \text{ otherwise} \end{cases}, \; j : t_j^{\langle i \rangle} \in [\tau_d, \tau_d + \Delta\tau) \tag{3.3}$$

Based on the prevailing sentiment of the posts, $o_d^{\langle i \rangle}$ takes values $\{-1, 0, 1\}$ with 0 indicating that user i did not publish any post during $[\tau_d, \tau_d + \Delta\tau)$, and $\{-1, 1\}$ standing for *Bearish* and *Bullish* opinions respectively. We further discuss details of *sentiment-* and *content-based* models as well as of couple of straightforward baselines.

Sentiment-Based

User's opinion can be thought of as a function of community sentiments. One objective of Stocktwits is to provide access to opinions and sentiments of active practitioners, so that one can use them as a basis for his own ideas. Indeed, if we narrow down the definition of community to users interested in a very small set of stocks (lets even limit it to a single company), we can see that it is not a very strong assumption. This model would actually do exactly the same as a real user does on a daily basis—process sentiments towards specific company and transform them into his own. In *sentiment-based* model we represent each individual opinion of a user i in a time bin d as an opinion vector of all other users during the same period: $U'_{id} = \{o_d^{\langle k \rangle}\}, k \in [1, K], k \neq i$. We then associate each U'_{id} with known opinion $o_d^{\langle i \rangle} \neq 0$ and train a separate model for each user i.

Content-Based

It is fair to say that different people might interpret very same event (fact) in various ways. Hence, previous approach is prone to relying on mistakes (or misinterpretations) of community members. Whereas, if a user had an access to original information (not second-hand), he could form a completely different opinion. Here, we assume that community content to some extent represents real-world events of potential interest for a target user. Thus, we believe that associating user's opinions with community-generated content instead of their distilled opinions would result in less bias. In *content-based* model opinion of a user i in time bin d is represented by a binary vector $U''_{id} = (w_1, w_2, \ldots, w_{|V_{\bar{i}}^{\cdot}|})$, where $V_{\bar{i}}^{\cdot}$ is global vocabulary of community:

$$V_{\bar{i}}^{\cdot} = \bigcup_{k,j,k \neq i} w_{M_{kj}}, j : t_j^{\langle k \rangle} \in [\tau_d, \tau_d + \Delta\tau) \ \forall d : o_d^{\langle i \rangle} \neq 0 \tag{3.4}$$

and $w_{r'}$ is defined as:

$$w_{r'} = \begin{cases} 1, \text{ if } \exists\ k, j \colon w_{r'} \in p_j^{\langle k \rangle} \text{ and } t_j^{\langle k \rangle} \in [\tau_d, \tau_d + \Delta\tau) \\ 0, \text{ otherwise} \end{cases} \tag{3.5}$$

Similarly to *sentiment-based* model, U''_{id} and corresponding $o_d^{\langle i \rangle} \neq 0$ are used to train a model for each user i. We restrain ourselves to unigrams for similar reasons as we did in *Bull-Bear* classifier. Admittedly, we did try to use topics model instead, however, latent Dirichlet allocation [13] was not able to distil separable topics. We attribute it partially to the fact that in this study we focused on Apple community only, and to shallow language used by practitioners. Temporal topics modeling also did not yield more comprehensive results, mostly because of simple reporting-anticipating patterns prevailing over related news and trends discussion during the whole timespan.

Baselines

We also propose couple of intuitive but not supervised approaches. *Majority vote* assigns $\widehat{o_d^{\langle i \rangle}}$ as an opinion prevailing in the community during $[\tau_d, \tau_d + \Delta\tau)$. *Historical activity* takes into consideration user i's overall disposition to express *bullish* opinions over *bearish*, or *vice versa*. Also since our dataset conforms to the fact that investors tend to share optimistic insights more frequently than pessimistic, even during the *bear* market, we examine the quality of *always bullish* and, to be consistent, *always bearish* baselines.

3.3.4 Individual Predictability

We do expect that performance of behavior prediction would vary from user to user, since some might be more active than others or have more homogeneous community

(read that community-generated content would be more representative of a user in such case). Therefore, it is crucial to know how different types of user profiles affect overall predictability. Such information would allow to assess a predictability of unseen users with a certain level of confidence. Recall that although opinion interpolation is interesting by itself, we pose this task as being a part of an outer predictive system. Thus, having derived opinions of low quality might (and definitely would) lead to error propagation and will degrade the quality of any model using such data. Alternatively, if confidence of prediction quality is known, one can keep only those interpolated opinions which do not negatively affect his system. We study characteristics of a user from three major perspectives that cover his content, network structure and level of engagement with the platform (see Table 3.2 for a full list of features).

Let f_i be F-measure yielded by a community-based inference model for a user i, and $pr_i = (ch_1^{\langle i \rangle}, ch_2^{\langle i \rangle}, \ldots, ch_{N_C}^{\langle i \rangle})$—his profile constituted by characteristics $ch_j^{\langle i \rangle}$, where N_C is a number of profile characteristics. We then associate each user profile pr_i with confidence of predictability $c_i = conf(f_i)$ derived from corresponding F-measure and fit a supervised model on $PR = (pr_1, pr_2, \ldots, pr_K)$ and $C = (c_1, c_2, \ldots, c_K)$.

3.3.5 Low-Rank Matrix Approximation

Problem we are tackling is well-known in image processing community as an approximation of missing or noisy pixels. State-of-art solution uses singular value decomposition (SVD) for low-rank approximation of a matrix representing target image. Major advancements come from different ways of initialization of the aforementioned missing elements. We set to assess whether this approach is a good fit for coping with absent opinions.

The goal is to populate all missing elements with either *bullish* or *bearish* opinions (*i.e.* a matrix depicted on Figure 3.1 has to be filled completely). However, in contrast

Table 3.2: List of activity, content and network features chosen to represent user profile

Activity
Number of active days
Number of overall posts
Number of daily posts
Content
Bull-Bear ratio
Average number of unigram
Vocabulary size
Similarity with community content
Usage of lexica popular within a community
Network
Number of followers
Number of following
Number of friends
Follower ratio
Number of mutual friends
All the above, but within the community

to settings normally used in computer vision community, where the original matrix with no corrupted elements is available, there is no "ground truth" opinion matrix in our case. Thus, we randomly withhold same percentage of known opinions for each user, initialize these values with labels assigned by different community-based inference models, apply low-rank approximation and compare a new matrix with original one. We provide the details below.

First, we detain some known opinions for each of the users, henceforth referred to as *missing elements* to be restored, from the original opinion matrix O. Then using the rest of available opinions we train separate models for every user to infer held out elements—new matrix $\widetilde{O}$ is obtained by initializing these *missing elements*. Then using SVD user-time opinion matrix $\widetilde{O}$ is factorized into $\widetilde{O} = U \times \Sigma \times V^T$, where U and V are matrices composed of left and right singular vectors of $\widetilde{O}$ correspondingly, and Σ is a diagonal matrix of singular values. We apply low-rank approximation by keeping only top r singular values, where r is a rank of the matrix $\widetilde{O}$ ($r \ll \{K, D\}$), and setting the rest of Σ to zeros. The resulting matrix $\widehat{O} = U \times \widehat{\Sigma} \times V^T$ is a low-rank approximation of $\widetilde{O}$ (and O correspondingly). Low-rank approximation changed the $\widetilde{O}$ matrix significantly, but since we are interested in approximating *missing elements* only, we introduce a matrix $\widehat{O}' = \{\widehat{o'_{id}}\}$, where

$$\widehat{o'_{id}} = \begin{cases} o_{id}, \text{ if } o_{id} \text{ was held out,} \\ \underset{l \in \{-1,1\}}{\operatorname{argmin}} |\widehat{o_{id}} - l|, \text{ otherwise} \end{cases} \tag{3.6}$$

because elements of $\widehat{O}$ are real-valued as opposed to discrete values stored in O (and $\widetilde{O}$). We then compare the disparity between original matrix O and matrices $\widehat{O}'$ obtained through different initialization of *missing elements*.

3.4 Experimental Results

3.4.1 Dataset Description

We crawled Stocktwits for consecutive period of 2013–2014 and limited posts only to those discussing top 10 most popular stocks on the platform (as defined by posting volume), with 8 of them representing hi-tech sector. Since the context of user opinions normally lies within events surrounding a single company, we defined communities based on the ticker of interest. Recall that all models introduced for community-based inference assume that we are dealing with active users determined to exploit the platform for their needs. Although there is a tremendous amount of "silent" users (about 70% of all), who can be either "listeners" or simply dead accounts, there is no way we can assess how they engage with the platform. Hence, we focus only on *active* users as defined by their contributions to ticker-related discussions. We select 200 most *active* users for each of 10 tickers based on the total amount of posts submitted during the observation period. However, for some of the companies (*e.g.* GOOG, NFLX and AMZN) even among top contributors there were some people who submitted less than 100 posts during 2 years. Although it can be explained by the fact that prominent Stocktwits users normally have a portfolio of several stocks that they monitor and report on, we are seeking for consistent activity with regards to each stock. Thus, we further subsample user pool to top 100 practitioners based on their posting frequency. Since AAPL showed noticeably denser activity matrix (see Figure 3.1) than other companies, we concentrate our analysis on this community, but it can be simply extended to others. Note that the intuition of proposed approaches is based on the idea that a user would take into account insights and/or sentiments of the most renowned peers in the community. Usually these are respected users dedicating significant amount of their time to constantly crafting their

Table 3.3: Statistics on the timelines of selected users in the pool

	Positive daily sentiment, %	Days with no content, %
min	21	19
avg	**59**	**55**
std	16	11
max	98	68

content. Although high level of activity cannot be a definitive indicator of correct and meaningful recommendations, we treat it as a proxy. Whereas, it is very difficult to judge about users with sporadic activity: even though their every comment and recommendation might be useful, other practitioners simply could miss them out, because they would not be that much visible in ticker streams as those very active contributors. We provide the details on the pool of 100 AAPL practitioners (total of 890K posts) in Table 3.3. This dataset is actually a very good example for the posed task, since on average every active user has more than half of his timeline "missing"[4]. Also note that our dataset conforms with standard tendency of oversharing bullish insights: although on average a typical timeline has a moderate dominance of bullish content (about 60% of time bins), for some of the individual timelines this ratio is extremely skewed.

3.4.2 Bull-Bear Classifier

We used 130K labeled posts subsampled from original dataset for *Bull-Bear* classifier. Its performance was evaluated using 10-fold cross-validation (see Table 3.4). Expectedly, lexicon-based approaches performed poorly, since they were able to assign a label definitively (however, not necessarily correct) only in 40% and 30% of cases for LIWC and financial lexicon respectively, while random sentiment was used for the

[4]There is no user-generated content during these time periods.

Table 3.4: Performance of Bull-Bear classifier yielded by lexicon-based and supervised models

	F_1
General	0.53
Financial	0.54
Supervised	**0.74**

rest of the posts. Supervised model (we used logistic regression as a classifier, since it has shown the best performance) yielded F-measure of 0.74. Interestingly, internal feature selector ranked highly terms that normally carry the highest semantic load in posts (*e.g.* short, bullish, divergence, higher, low, *negative emoticons*, inverted, holding, bought, *etc.*). Thus, we can conclude that lexicon-based approaches are not sufficient for the sentiment analysis task when applied to short segments of financial text.

3.4.3 Community-Based Inference

We conducted experiments on both 100 and 200 active AAPL users, however, since obtained results were comparable, we report only those for 100 users. Recall that in community-based inference we build a separate model for each target user (see Section 3.3). The obvious way to assess the performance is to do that on a *user-level* and then average it over all users from the pool. This is very informative when one is interested in subsampling the pool by keeping those users who exhibit predictable behavior. However, a typical task would be simply to populate all missing entries of original opinion matrix (see Figure 3.1). In such case decent *user-level* performance is not as crucial as *opinion-level*. Thus we provide analyses for both macro and micro levels. In either case we employ 10-fold cross-validation applied to each target user.

Table 3.5: User-level F-measure yielded by classifiers employed in *sentiment-* and *content-based* models

	Sentiment-based	Content-based
SVM	0.52	**0.53**
Logit	0.52	0.52
Naive Bayes	**0.57**	0.45
Random Forest	0.52	0.51

Table 3.6: User-level performance yielded by sentiment- and content-based models, majority vote and historical activity baselines. F_1^- and F_1^+ stand for F-measure of negative and positive classes respectively.

Model	F_1	F_1^-	F_1^+
Sentiment-based	**0.57**	0.49	0.65
Content-based	**0.53**	0.40	0.66
Majority vote	0.47	**0.25**	0.67
Historical activity	0.50	**0.38**	0.62
Random guess	0.47	0.41	0.53
Always bullish	0.38	**0.02**	0.75
Always bearish	0.27	0.53	**0.01**

User-level performance

We tried different classifiers in application to two supervised models proposed—*sentiment-* and *content-based* (results can be seen in Table 3.5). We further use naïve Bayes for *sentiment-based* and support vector machine (SVM) for *content-based* models. As it can be seen from Table 3.6, neither *majority vote* nor *historical activity* models were able to cope with extreme class imbalance. Surprisingly, assigning *always bullish* sentiment did not result in high F-measure as well. Both supervised models performed better than biased random (*historical activity*) with *sentiment-based* yielding slightly higher F-measure than *content-based*.

Opinion-level performance

We can see better performance for all non-constant models when assessing on *opinion-level* (see Figure 3.3): *content-* and *sentiment-based* yielded F-measure of 0.63 and 0.61 correspondingly, with *historical activity*, and *majority vote* also performing better ($F_1 = 0.56$ and $F_1 = 0.51$ respectively). Also, in contrast to *user-level* evaluation, *content-based* outperformed *sentiment-based* model with significantly bigger area under the curve (AUC=0.7).

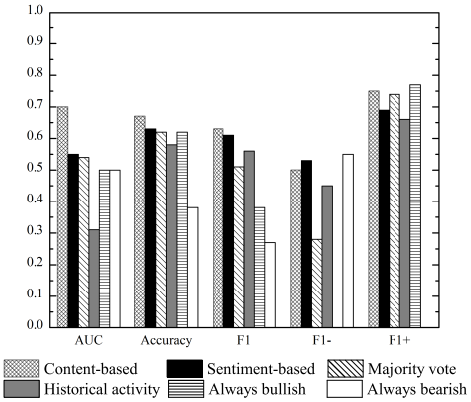

Figure 3.3: Opinion-level performance yielded by sentiment- and content-based models, majority vote and historical activity baselines. F_1^- and F_1^+ stand for F-measure of negative and positive classes respectively.

Hence, we examine receiver operating characteristic (ROC) curves for both supervised models on Figure 3.4. To make sure that such result was not obtained because of *content-based* model favoring majority class (*bullish* opinions), we compare ROC curves for both *bullish* and *bearish* opinions. As it can be seen, *content-based* model dominates on both supporting our hypothesis that on average it carries less user bias than *sentiment-based* model.

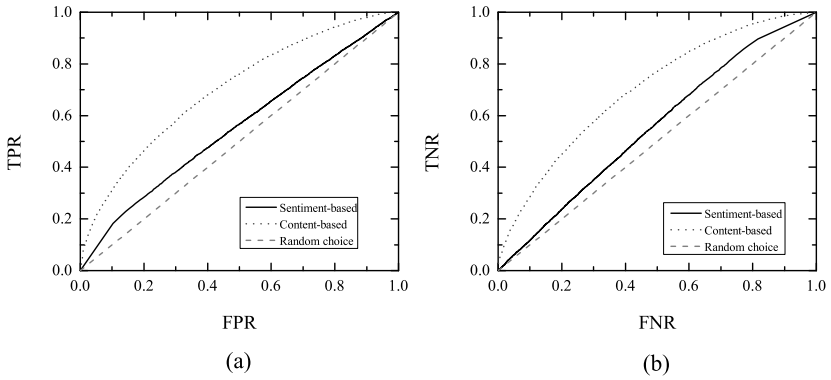

Figure 3.4: ROC curve of opinion-level performance when inferring *bullish* (a) and *bearish* (b) opinions. *Random choice* also resembles ROC for *always bullish* and *always bearish* baselines.

3.4.4 Individual Predictability

Let us return back to *user-level* evaluation and examine performance of *sentiment-based* model in details. Despite the fact that average F-measure was 0.57, more than 40% of users exceeded this value on individual level (see Figure 3.5). Hence, it is indeed crucial to differentiate between users exhibiting high and low predictability.

We would like to make one more observation before moving to the model assessing dependency of predictability on different types of user profile. Recall that we evaluated individual predictability using 10-fold cross-validation. We discovered that highly predictable users have an excessive variance in F-measure across 10 folds. Indeed, individual predictability and cross-fold variance have a very strong correlation of 0.54 (see Figure 3.6). Then we can speculate that users which tend to reflect on relatively small changes of a stock price are more likely to be correctly predicted based on their community. We also believe that variance in cross-fold predictability can be used as a low/high-predictability filter by itself.

For the model estimating confidence of user's predictability, we experimented with

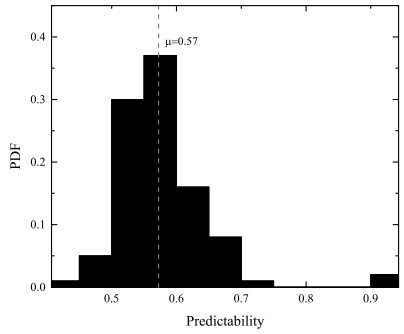

Figure 3.5: Probability density function of individual predictability yielded by *sentiment-based* model. Although F-measure averaged over 100 users is equal to 0.57 only, almost 40% of users exhibited individual predictability above this value.

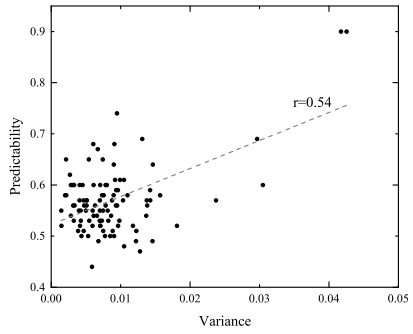

Figure 3.6: Relationship between variance in individual F-measure across k folds and overall user's predictability. Pearson's $r = 0.54$.

$c_i = conf(f_i)$ to be real-valued, discrete and conclude that the best results were achieved when deciding on user's predictability below and above $f_{th} = 0.5$. Thus, we cast this problem as a binary classification of predictable and unpredictable users and report the results yielded by random forest model (as the one showing superior performance) on different sets of profile features (see Figure 3.7). Best performance (F_1=0.65) was achieved by set of content-based features solely, with additional user characteristics degrading the results. It is not surprising, as we can see that random

Table 3.7: Importance of individual features (content and activity only) as defined by Random Forest classifier

Feature	Importance
Bull-Bear ratio	**0.24**
Active days no.	0.12
Vocabulary size	0.12
Density of community content feature matrix	0.11
Overall posts no.	0.11
Content similarity overall	0.07
Daily posts no.	0.07
Content similarity projected to user's lexicon	0.06
Content similarity daily	0.04

forest assigns the highest importance to *bull-bear ratio* (see Table 3.7) as the feature informative of class imbalance for individual users. Classifier also highly ranks the *number of active days* and *size of individual vocabulary*. Based on the experimental results we believe that both profile-based and variance-based models can be used to filter out unpredictable users.

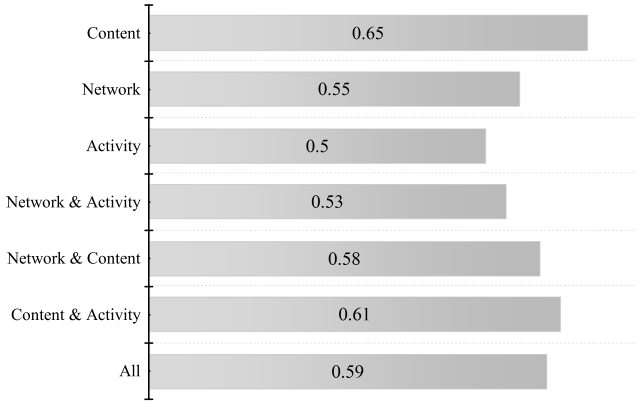

Figure 3.7: F-measure yielded by different feature sets for user predictability model

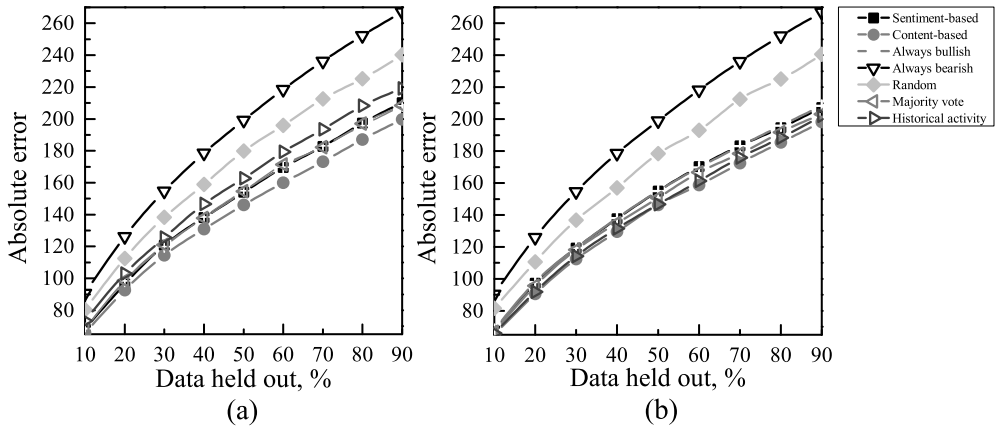

Figure 3.8: Matrix disparity before (a) and after (b) low-rank approximation (r=2) for different amount of data being held out

3.4.5 Low-Rank Matrix Approximation

Since we are working with opinion matrix, we expect the results of low-rank approximation closely resemble performance reported in *opinion-level* assessment. We analyze how disparity between original and approximated matrices changes with different amount of elements being held out. We also perform a sanity check by analyzing the magnitude of disparity between original and initialized matrices O and $\widetilde{O}$ to see if approximation results in any improvement compared to simple filling of missing elements with community-based models. As can be seen from Figure 3.8b, the minimal error was obtained by *content-based* model closely followed by *historical activity*, *majority vote* and, surprisingly, *always bullish*. *Sentiment-based* performed worse than aforementioned models. Next, we examine the magnitude of absolute error yielded by champion models when comparing original matrix with initialized and approximated ones (see Figure 3.9). Only three leading models showed the improvement brought by low-rank approximation, with *historical activity* decreasing the error by 10% on average. *Content-based* model was superior to others both before and after matrix approximation and was able to achieve an improvement in disparity of about 1% after

SVD was applied.

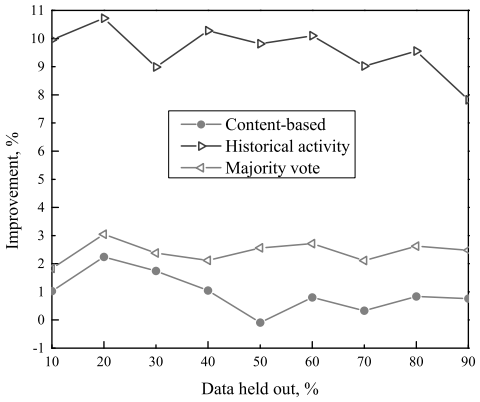

Figure 3.9: Improvement in matrix disparity when comparing original matrix with initialized and approximated matrices accordingly

3.5 Conclusion

Summary

In this chapter, we studied the possibility of predicting missing opinions of users which have already had a moderate activity in a microblog. We proposed a number of models that interpolate one's absent opinions based on the content generated by the community and his previous activities online. We have shown that both supervised models we introduced—*sentiment-* and *content-based*—outperform intuitive baselines and achieve decent performance when assessed both on *user* and *opinion* levels respectively. We have also shown that state-of-the-art technique for matrix restoration improves the quality of opinion inference if initialized using proposed models. We analyzed how behavior predictability varies across users and have shown that a model can be built to discard opinions of highly unpredictable users either based on characteristics of the content they author and engagement with the platform or the variance

in the cross-fold predictability, which can be easily gauged on validation set.

Implications

The techniques we introduced are not limited to the stock market domain. They can be effectively applied to other disciplines where users share their recommendations with respect to one or more entities of interest. This work is aimed to propose solutions to mitigate challenge with missing data when dealing with user-centric techniques. Speaking of user-centric analytics, one can argue that these approaches might violate users' privacy. While this issue is an ongoing discussion in social media community, we suggest to take into consideration the following angles: (i) in all models, we do our due diligence to protect users' personal information by appropriate and effective anonimization; (ii) we work only with user's opinions which were publicly posted; (iii) even in user-centric analytics, proposed techniques concentrate on collective behavior rather than on individual one; (iv) and finally, in any circumstances, user is not and should not be liable to what one infers as his missing opinion.

Future work

There are many ways in which this research can be continued. Although the choice of community inherently implies some level of homogeneity, not all members are equally important and representative of each target user. Thus, it would be interesting to implement a filtering mechanism based on social relevance which would learn one's opinion from informative content only. Also in this study we concentrated only on investment community, while if applied to a broader group of people which are not united by a single motif (*e.g.* arbitrary Twitter users), different strategies might be required.

Chapter 4

Social Filtering

4.1 Introduction

More and more tasks nowadays become crowdsouced. Starting with community-powered encyclopedias, goods and services reviews, data processing and annotation up to content moderation and detection of sybils. Some are delegated to crowdworkers, others require volunteers' participation, and the rest completely rely on the wisdom of crowds. Quite often such tasks involve some sort of content generation—it can be as simple as labeling/rating or more time-consuming, such as text summarization or writing an essay. Naturally one would like to know which of the aforementioned methods produces the content of a higher quality. Crowdworking marketplaces, such as Amazon Mechanical Turk[1] and Crowdflower[2], gained popularity recently because they allow to distribute bulky tasks among a large number of workers for relatively cheap price. However, since the jobs are usually routine but time- and attention consuming, some workers tend to perform more jobs in order to get higher profit by sacrificing the quality [54, 68]. Voluntary participation, also known as collaborative

[1] `mturk.com/mturk/welcome` (2015-07-01)
[2] `crowdflower.com` (2015-07-01)

filtering, does not involve financial motivation, and so it is fueled by pure enthusiasm and one's will to help the community. There are three drawbacks with this approach though. First, since there are no obligations, it is not clear whether the certain amount of content would be produced, and, if so, when it will happen. Next, humans tend to change their behavior if they are aware of the fact that they are being observed. For instance, if person believes that there are certain expectations from him, then he would try to meet them rather than expressing thoughts honestly (*e.g.* imagine a study where immediately after an exam students are asked to anonymously answer whether they were cheating or not). Finally, even in a casual environment (*e.g.* Facebook) people frequently censor themselves in order to avoid confrontation with others [32, 84].

However, in a more relaxed setting where a platform carries less characteristics of an offline community (*e.g.* Twitter as opposed to Facebook), users are way more eager in expressing themselves. No matter if people are considering their imaginary audiences before publishing tweets or not, they report to post genuine and emotional content [67]. Considering that, there is a plethora of information that is highly characteristic of individuals that authored it. In this work, we introduce a concept of *social filtering* as a new approach to collaborative problem solving. In contrast with previously discussed methods, it leverages content published earlier (see Figure 4.1 for a comparison of approaches). Its objective is to determine a subset of individuals highly suitable for a chosen task and then to derive the most informative features from the posts they submitted. Therefore, *social filtering* takes advantage of the rich amount of publicly available data ready to be used. The method invokes a notion of *expert* content as opposed to data aggregated across all users. According to Zafar *et. al.* [109], such strategy outperforms random sampling in a number of facets, including popularity, topical and opinion diversity, timeliness and trustworthiness.

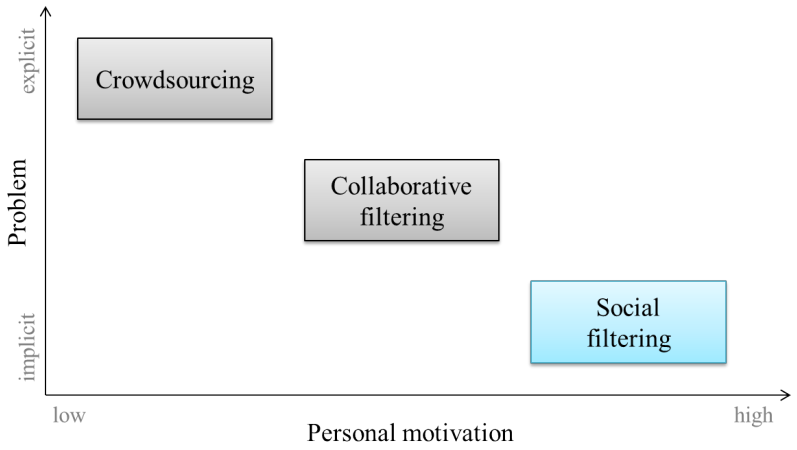

Figure 4.1: Collaborative problem solving with regards to content generation. By *personal motivation* we mean one's enthusiasm and desire to publish some content that is not fostered by external factors.

The method is evaluated on a case study of crime trend prediction for Chicago, IL, which is the third largest city of the US [2], and featured in 2013 FBI report [1] as top one in murders, top two in robbery and top three in property crimes. There is a number of underlying factors affecting future crimes, such as unemployment and divorce rates, education level and average income, community welfare, general happiness of the population, and many others. We hypothesize that by having a *"perfect sample"* of *"expert users"*, in our case arbitrary citizens active on Twitter, we would be able to extract those hidden variables and translate them into predictive signals.

The remainder of the chapter is organized as follows. We review previous works on crime prediction in Section 4.2 and describe proposed model in Section 3.3. Details on a choice of datasets, user representation and expert selection as well as experimental results are presented in Section 3.4. Finally, in Section 3.5 we discuss the implications of proposed approach and directions of future work.

4.2 Previous Work

Social filtering is an approach which is designed to handle a variety of different tasks that fit into a scope of trend and event prediction based on user-generated content. However, selection of *expert* users is unique to every particular problem, since the definition of individual's *relevance* and *value* highly depends on a context. Hence, based on a choice of the case study we only review the works concerned with crime prediction.

Various approaches were undertaken with respect to this problem. Conventional techniques used by law enforcement agencies used location-centric paradigm and were predominantly based on density functions for generating hotspot maps [19,36]. However, they are peculiar to particular location, thus cannot be generalized. To overcome this issue subsequent methods aimed to incorporate background knowledge about geographical features of a region, such as distance to intersections and highways, schools and businesses, parks and hospitals, demographics of the neighborhood as well as other information [97,104].

User-centric techniques extensively explored various socio-economic factors, such as level of education, average income, unemployment rate, female-to-male distribution, racial background and many other [16, 29, 37, 39, 70, 75]. Social structure of a community, which was considered to be a key factor controlling criminal activity, was also studied as a part of user-focused approach [47,88]. Behavioral-based models examined predictive power of mobile network activity [14,89]. Different line of research modeled future crimes as a self-exciting point process [25,71,86].

The closest works to ours are by Wang *et. al.* [98] and by Gerber [40]. The former was the first to blend social media data into traditional models. Authors predicted city-wide hit-and-run accidents using event-based topics extracted from local Twitter

accounts. However, tweets were limited to those published by a set of manually selected news agencies, this way vast amounts of information produced by regular users were ignored. Also underlying assumption that news are relevant to the recent events occurred in the town is not necessarily correct for all cases. Finally, it is not clear whether the model is generalizable to other crime types. Gerber proposed a model extending kernel density estimation (KDE) with Twitter-derived topics. Although content fusion has introduced some additional context, KDE is based on geospatial records, hence it lacks portability and cannot be directly applied to other cities. KDE also does not take into account temporal changes in criminal activity. Another close work [92] used social media to observe whether criminal activity has an effect on public sentiment. Authors have shown that long-term historical incidents correlate with elevated negative emotions of respective local community on Twitter.

Current study is based on our previous paper [23]. Unlike reviewed works, instead of using content aggregated city-wide, we focus on individual users which allows to select the most informative ones and also to discard the noise. Besides, it is location-agnostic, since it only depends on Twitter data.

4.3 Our Approach

Let $T_i = \{\left(p_1^{\langle i \rangle}, t_1^{\langle i \rangle}\right), \left(p_2^{\langle i \rangle}, t_2^{\langle i \rangle}\right), \ldots, \left(p_{N_i}^{\langle i \rangle}, t_{N_i}^{\langle i \rangle}\right)\}$ denote a timeline of a user i, where tuple $\left(p_j^{\langle i \rangle}, t_j^{\langle i \rangle}\right)$ represents user i's post j along with its timestamp: $t_1^{\langle i \rangle} < t_2^{\langle i \rangle} < \ldots < t_{N_i}^{\langle i \rangle}$. Post $p_j^{\langle i \rangle} = \{w_1^{\langle i \rangle}, w_2^{\langle i \rangle}, \ldots, w_{N_j^{\langle i \rangle}}^{\langle i \rangle}\}, N_j^{\langle i \rangle} \in [1, |V|]$, is comprised of tokens $w_{jk}^{\langle i \rangle}$. $V = \bigcup_{i,j,k} w_{jk}^{\langle i \rangle}$ is a global vocabulary, $k \in [1, N_j^{\langle i \rangle}]$.

Let $\Delta\tau$ be the smallest unit of aggregation with $(\tau_1, \tau_2, \ldots, \tau_D), \tau_d = \tau_{d-1} + \Delta\tau$, denoting edges of a sequence of consecutive time bins. User i is represented by $U_i = (U_{i1}, U_{i2}, \ldots, U_{iD})$, where D is a total number of time bins across all users in

the pool. Then $U_{id} = (v_{d1}^{\langle i \rangle}, v_{d2}^{\langle i \rangle}, \ldots, v_{dM}^{\langle i \rangle})$, where v_{dm} is a function of all tweets of user i from time bin d:

$$v_{dm}^{\langle i \rangle} = f_m(w_{kj}^{\langle i \rangle}), j : t_j^{\langle i \rangle} \in [\tau_d, \tau_d + \Delta\tau) \tag{4.1}$$

M specifies a number of features for selected representation. Thus the dataset is defined and consists of D samples $U_d = (U_{1d}, U_{2d}, \ldots, U_{Nd})$.

Let a_d be a number of incidents reported during $[\tau_d, \tau_d + \Delta\tau)$. Then a label l_d for bin d is defined as:

$$l_d = sign(a_d - a_{d-1}) \tag{4.2}$$

Model is trained on a set of $D - l$ samples $U = (U_1, U_2, \ldots, U_{D-l})$ and corresponding labels $L = (l_{1+l}, l_{2+l}, \ldots, l_D)$, where $l \in \mathbb{Z}$ is a lead or lag between content and incidents.

4.4 Experimental Results

We start by describing our datasets and proceed with a discussion on the choice of user representation. We show that there is a need in an appropriate procedure for selection of *expert* users, hence three-step filtering technique is introduced and evaluated. Then, inspired by [92], we examine if users' tweeting behavior is indeed predictive of future incidents or it is rather "reporting" on current crimes. We also study to which extent model is dependent on both new and historical data. And finally, observe how quality of prediction changes across different crime types.

We set $\Delta\tau$ to a single-day resolution for all of the following experiments. Movements of crime trend are defined simply as a sign of difference between two consecutive

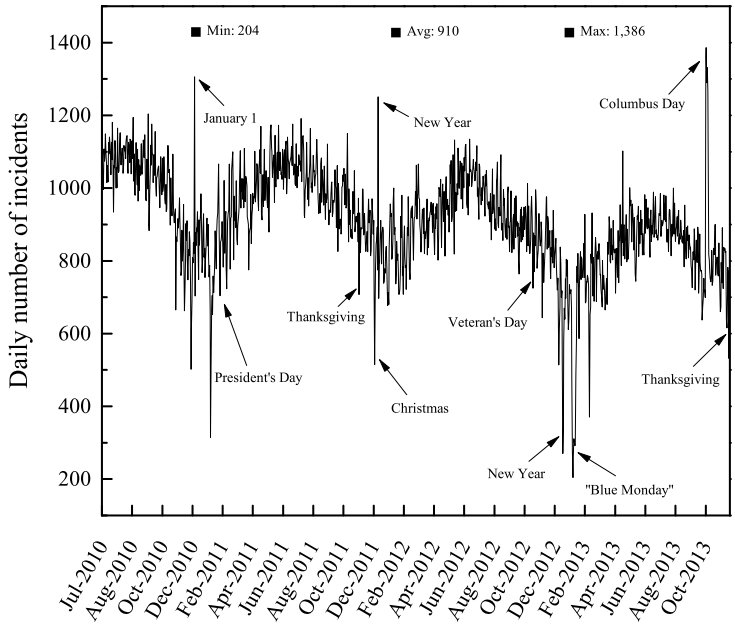

Figure 4.2: Daily aggregated indicents. Spikes were observed during statutory holidays.

days[3]. Here, we focus on a near-future forecasts with lags set to $l \in [1, 14]$.

4.4.1 Dataset Description

Chicago Police Reports

Our dataset was collected from official Data Portal of Chicago[4] and covers time range between July 1, 2010 and November 30, 2013 with a total of 1.7M incidents. Every instance is reported together with its exact location (both longitude-latitude pair and full address), timestamp and crime type (see Table 4.1 for details). Figure 4.2 shows that time series has a strong periodical nature, which conforms with seminal

[3]Strategies accounting for both magnitude of the change and actual number of incidents were also tested, however, they showed weaker performance.

[4]`data.cityofchicago.org/Public-Safety/Crimes-Mapdfnk-7re6` (2015-07-06)

works in criminology community [6], [57]. The overall downtrend that started in US in 1990s [10] can also be seen on this figure. Extremes with remarkably low or high criminal activity occur on national holidays, however, whether particular one would result in rise or decline cannot be determined definitively (see New Year, for example). Although such outliers and daily fluctuations could be removed by smoothing, we refrain to do so, since in that case we would not be modelling actual crime trend.

Twitter set

Corresponding set of Twitter users from Chicago was collected based on Online Coupling from the Past [100]. It guarantees the convergence on a "perfect sample" of the whole user network while being unbiased towards individuals with extreme number of connections. Historical timelines[5] of selected users were retrieved and restricted to the same timeframe—between July 1, 2010 and November 30, 2013. Daily statistics on posting volume are depicted on Figure 4.3. However, because of Twitter rate limits, retrieving historical posts even of the "perfect sample" results in *active* users being represented only by their recent tweets, while the timelines of people with low or moderate levels of engagement reach far back in the past. These challenges should be addressed when developing a method robust to missing and inconsistent data. We report per-user activity statistics in Table 4.2.

4.4.2 User Representation

We posit that representative sample of citizens can reflect emotional and moral state of the whole society. Hence we choose LIWC [77]—widely adopted psycholinguistic lexicon—as a mean of deriving user-dependent features. We extract the following measures of affect: *positiveness*, *negativity*, *sadness*, *anger* and *anxiety*; as well as a

[5]We would like to thank Kenton White for providing access to this dataset.

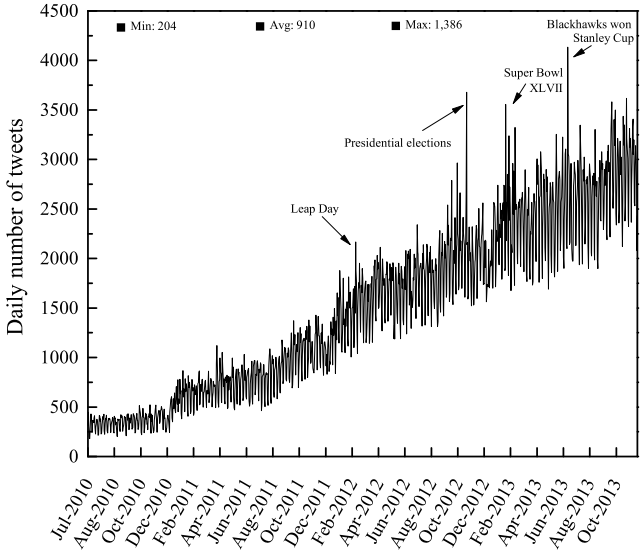

Figure 4.3: Daily tweet volume. Peaks are labeled with corresponding topics trending in our set during those days.

fraction of *obscene* and *death*-related words.

User is then defined by each of the aforementioned metrics solely as well as by their combination. Also, similarly to [23], we examine discrete representation of the affect (positive, negative or neutral). Results are presented on Figure 4.4. Surprisingly, *positive affect* alone achieves the best performance ($F_1 = 0.59$)—even better than representation comprised of all 7 features. We expected metrics corresponding to negative emotional expression to be more informative about future crimes, as opposed to public positive sentiment. Thus to further investigate this matter we plot daily *positive affect* aggregated across all users along with corresponding number of incidents (see Figure 4.5).

It can be seen that long-term decrease in criminal activity is reflected by a moderate but steady uptrend in *positive affect* of a user population. Also, all significant

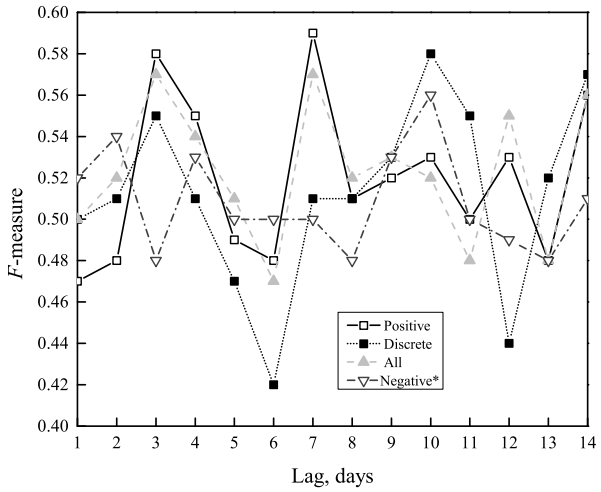

Figure 4.4: Performance of user representation for varying lag. *Negative*, *sad* and *anxious* affects as well as fractions of *obscene* and *death*-related words exhibit the same trend, hence we report only *negative affect*.

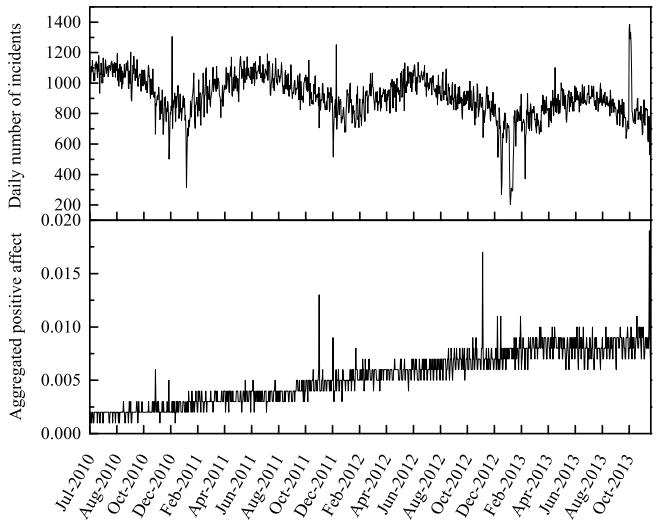

Figure 4.5: Daily number of incidents with corresponding positive affect aggregated across all users

drops in number of incidents are preceded by spikes in *affect*[6]. The latter is explained by mass anticipation of upcoming holidays. One then might argue that *experts'* content can be replaced by a flag indicating whether day d is a statutory holiday. However, such approach has the following drawbacks. First, effect caused on a crime trend direction is not consistent even between exactly same events, and from year to year it can lead to the opposite consequences (see Figure 4.2). Second, while US celebrates only 10 federal holidays, it is not clear how the movement should be inferred for a regular day. Finally, such notation does not capture the global decline in criminal activity. We conduct similar analyses on other measures of affect, however, such pattern is not repeated. Therefore in all subsequent experiments we adopt *positive affect* to represent a user.

4.4.3 Experts Selection

Although the purpose of the *perfect sampling* is to select the best set of users (in our case specified as users from Chicago), its notion of "best" does not have to align with application-specific definition. That said, not every user is equally valuable for the problem of crime trend prediction. The purpose of *social filtering* is to automatically identify those individuals whose content is highly informative about future crimes. Here, we introduce a *three-step filtering* procedure including *activity-* and *relevance-*based filtering and *ensemble* of expert users.

Activity-Based Filtering

Twitter API allows to access only 3,200 most recent posts of its users. This results in highly active individuals dominating the latter part of the dataset while being generally underrepresented across the whole timespan (see Figure 4.6a). Another

[6]Also note that this lag seems to be constant for all events.

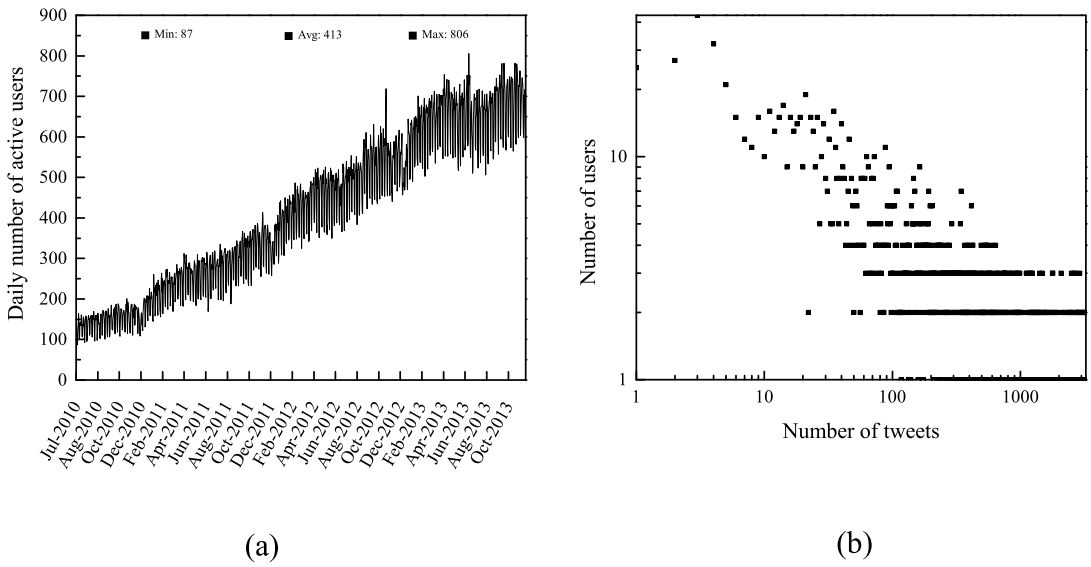

Figure 4.6: User activity in terms of days with published content and total number of posts. (a) Daily number of unique active users. Total number of users between July 1, 2010 and November 30, 2013 is 2753. On average only 15% of all users contribute to daily content. (b) Distribution of posts between users. Log-log scale is used.

issue introduced by Online Coupling from the Past is that it returns a subsample of nodes with the same characteristics as in original network, including the ratio of users with low levels of engagement. Indeed, generation of content in our sample exhibits a power law distribution (see Figure 4.6b) which conforms to statistics reported for online social networks. This means that the vast majority of tweets in our dataset are authored by a small number of users. Which subsequently leads to a question: how much value is added by a user producing extremely small amount of content (for instance, 1 post—see Table 4.2)?

Hence, first step is concerned with user filtering based on their *activity*. We apply radix sort to rank them based on the number of days with reported activity and total number of posts. Performance for varying number of selected users is presented on Figure 4.7a. Better results are consistently observed on $l = 7, 14$ with the best performance ($F_1 = 0.62$) occurring on 500 users with the lag of one week. We next

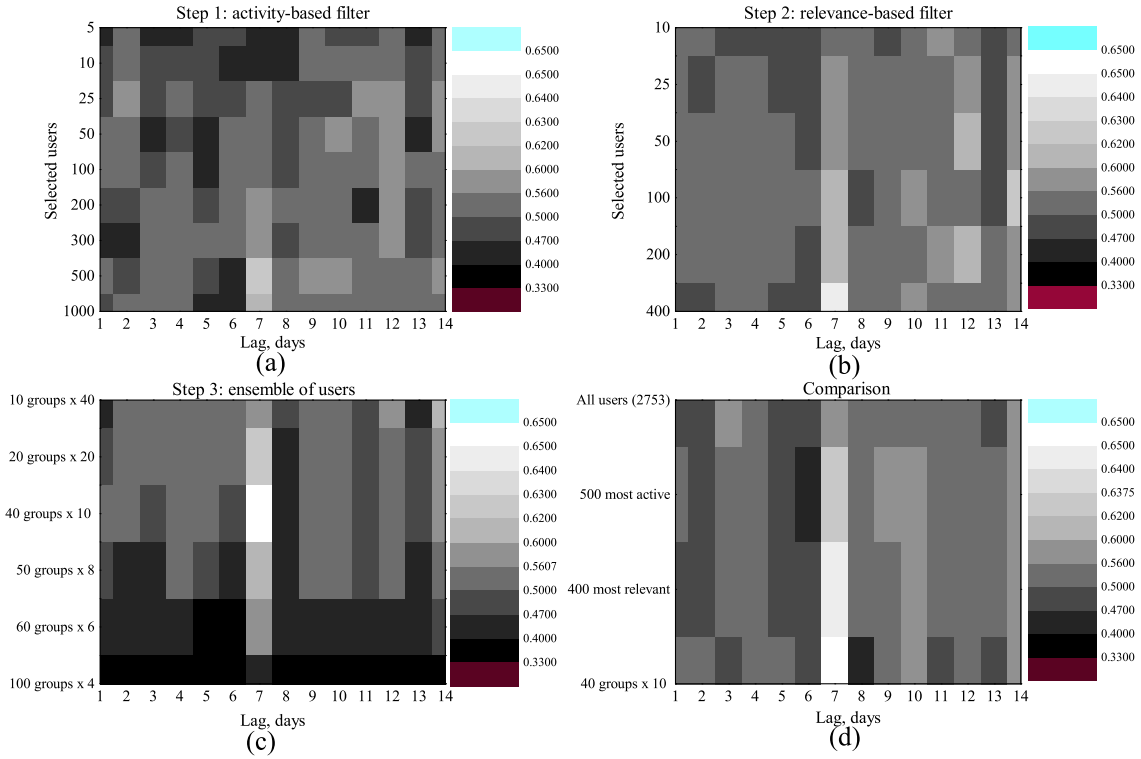

Figure 4.7: (a-c) F-measure obtained during each phase of three-step filtering. Rows correspond to single experiments. Lighter colors stand for higher values. (d) Comparison of the best results: (i) before filtering (all 2753 users are kept), (ii) for activity-based, (iii) relevance-based filters and (iv) ensemble of users.

discuss how the predictability can be further improved using *relevance*-based filtering.

Relevance-Based Filtering

Although previous step discarded people having low contributions, we are not actually interested in all active users. Instead, the goal is to narrow down this subsample even more—to keep only those individuals whose sentiment (*positive affect*) is highly correlated with the movements of crime trend. In this phase we consider top 500 users selected by *activity*-based filter as initial pool, and eliminate *irrelevant* ones based

on their χ^2 score [105]. We report results on Figure 4.7b. Compared to the previous step, 7-day trend is more noticeable with results for any number of users also being better. The best performance is observed for 400 users ($F_1 = 0.64$) decaying slowly with number of people decreased to 10. Now we introduce *ensemble of experts* aimed to reduce confusion of the model.

Ensemble of Experts

Objective of *social filtering* is to process content that is produced by people without understanding the purpose it is going to be used for. No matter how general and applicable it sounds (recall that it does not solicit for human contributions), it has its own shortcomings. There has to be a unique transformation of user-generated content to feature space used for modeling. The latter is designed based on assumption which holds for majority of individuals in the *expert* sample, but not necessarily for all of them. Thus forcing such transformation results in inconsistent features for some of the users, which, in turn, increases perplexity of the model.

To alleviate this issue we next employ an *ensemble* of supervised models using vertical partitioning of the U, that is each model is fitted only on a subset of *experts*. We use a non-overlapping partitioning based on *experts'* scores. Performance obtained for different grouping settings is shown on Figure 4.7c. Similarly to previous experiments, 7- as well as 14-day trends are clear. The best predictability is attained for 40 groups of 10 experts ($F_1 = 0.65$). Comparison of F-measure yielded by each step as well as by a baseline including all users is presented on Figure 4.7d. It can be clearly seen that each step introduces additional improvement. Therefore we apply *three-step filter* to all of the subsequent experiments.

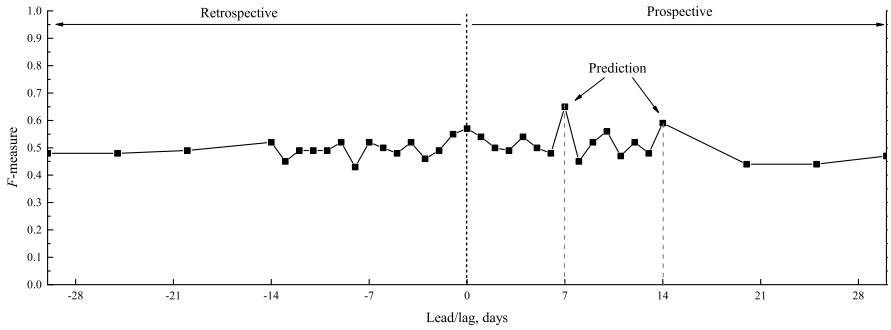

Figure 4.8: Predictability observed for different leads and lags. Areas corresponding to *reporting* and *predicting* behavior lie to the left and right of $l = 0$ respectively. Best predictability occurs on $l = 7$ and $l = 14$.

4.4.4 Prediction *vs.* Reporting

While earlier we speculated about the ability of posts published by arbitrary citizens to give some insights about prospective criminal activity, we did not discuss how this content relates to historical or contemporary incidents. Although we do not claim the causality between these two, comparable correlation can be observed for an opposite direction. Hence, similarly to Valdes *et. al.* [92], who used analogous LIWC-based features derived from aggregated content, we would like to assess the presence, and if the case, the magnitude of so-called *reporting behavior*.

We plot prediction quality for both leads and lags between 1 to 30 days (see Figure 4.8). First, we would like to note that performance of *reporting* behavior (*retrospective* part of the Figure 4.8) is comparable to as of a random guess. Second, *prospective* part shows significantly higher performance with its peaks on 7 and 14 days ($F_1 = 0.65$ and $F_1 = 0.59$ respectively), as observed earlier. After that F-measure gradually decays ($F_1 = 0.47$ for $l = 30$, which is worse than random guessing).

Our results conform with [92], even though the authors targeted a different location, analyzed bulk content and concentrated on exact-value prediction of emotional expression. Thus we can conclude that tweets published by *expert* users are rather

predictive of future crimes than reporting on historical (or current).

4.4.5 Dependency on Historical Data

As can be seen from Figure 4.2, Chicago is not exceptional in a sense that its crime trend suffers strong seasonality [6, 57]. And since our formulation of the problem is that we are looking to predict a binary event—rise or fall of criminal activity—completely neglecting its magnitude, it is valid to check whether older data corresponding to the same season exhibits same predictability.

With this question in mind we conduct two experiments to see, first, whether model fitted on old data is capable of forecasting directions for timespan which is further in the future, and if not, what the maximum lag between training and test data that guarantees a decent performance is; and, second, what amount of historical data is needed to achieve a satisfactory quality of prediction.

Experiment setting is shown on Figure 4.9. Test set is fixed between April 1, 2013 and November 30, 2013, in both scenarios. The only difference between two is that of direction in which training set is being increased: in first scenario (see Figure 4.9a) it starts in the beginning of the dataset—July 1, 2010—and progresses towards the test set, while second (Figure 4.9b) uses the opposite direction.

We report our results on Figure 4.10. When predicting with older data (see Figure 4.10a), it can be seen that model is able to achieve a substantial performance ($F_1 = 0.6$) only after 61% of the data is observed, which means that training set has to be no older than 8 months compared to the test. Another conclusion that we can derive here is that since our results do not exhibit seasonal dependence, content-based supervised model cannot be simply replaced by a time series exploiting historical trend. As for dependency on historical content (see Figure 4.10b), when dataset is up-to-date, a decent performance ($F_1 = 0.6$) is attained even with training set

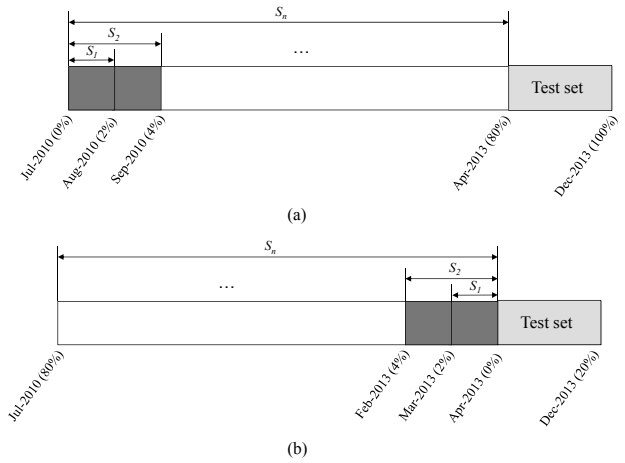

Figure 4.9: Experiment scenarios. Test set is equal to 20% of the dataset and is fixed between April 1, 2013 and November 30, 2013. Scenario 1 (a) uses training set starting in July 1, 2010 and progressing towards test set, while for scenario 2 (b) starting point is a beginning of a test set and it increases in retrospective direction.

constituting only 12% of the available data. This implies that *experts'* content, to be precise their individual *positive affect*, is able to capture the essence of current events, which are obviously transient—hence the low performance of the first scenario even when trained on a large fraction of the dataset (Figure 4.10a).

4.4.6 Predictability by Type

We have shown that prospective crime trend can be forecasted using positive affect of the *expert* citizens. However, all previous experiments were concerned with the aggregated number of incidents. Here, we examine whether the same pattern holds for various categories of the crime.

We repeat the same experiment on a finer-grain level of individual types with that only difference that we also allow for same-day prediction, as we observed on

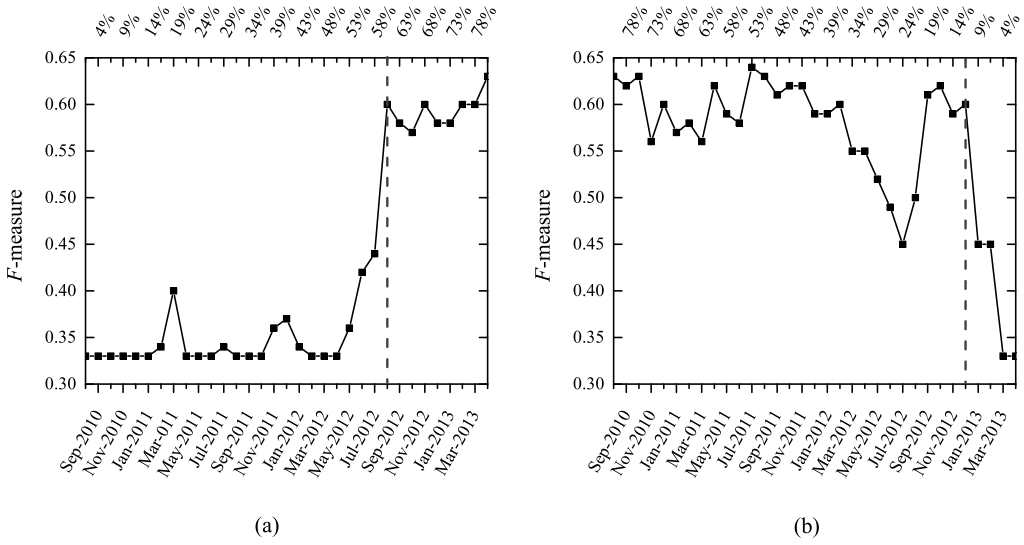

Figure 4.10: Model performance for scenarios: (a) classifier deterioration and (b) dependency on historical data. Top axes identify the size of corresponding training set.

aggregated data that it also offers satisfactory performance ($F_1 = 0.57$), yet slightly worse if compared to $l = 7, 14$.

F-measure and area under the ROC curve (AUC) are reported for 5 best-performing categories along with results for total number of incidents (see Figure 4.11). While also having peaks on $l = 7, 14$ or around ($l = 6, 13$ for some), most of the categories show best performance for the same-day prediction scenario (*e.g.* $F_1 = 0.63$ for burglary). Prediction quality of other types on average is no better than as of a random guess. While it can be attributed to a relatively small number of incidents per category (see Table 4.1)—presented top 5 categories contribute to 36% of all instances—, we can only speculate about other reasons (*e.g.* battery, which is top 2 category according to the total number of reported incidents, cannot be successfully predicted).

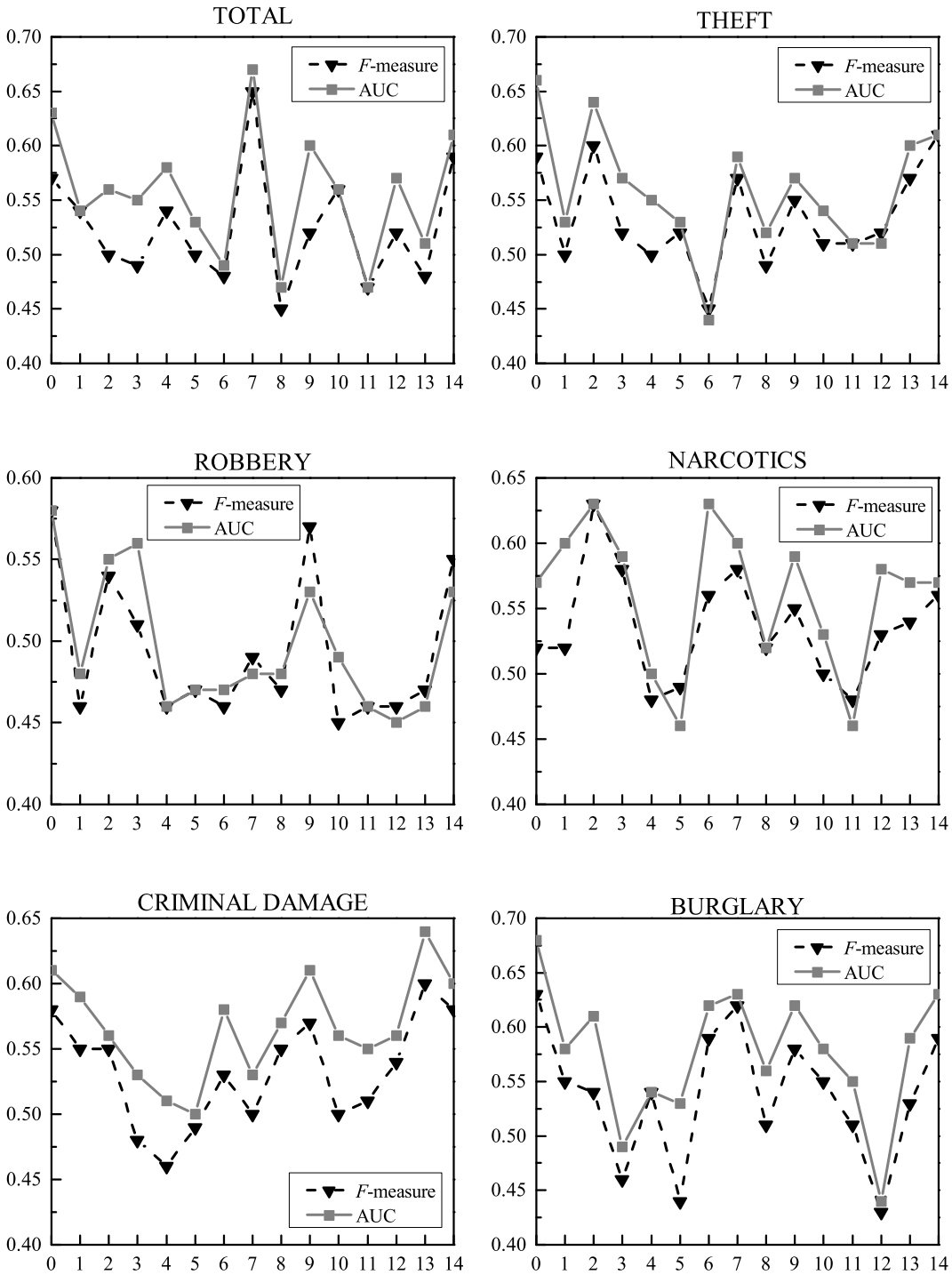

Figure 4.11: F-measure and AUC of the model for crime types which yielded the best performance. Lags (in days) are specified on x-axis. $l = 0$ stands for the same-day prediction.

4.5 Conclusion

Summary

We presented *social filtering*—a novel user-centric approach for predicting social trends. Its main advantage lies in the fact that it leverages authentic posts deliberately published by users without overthinking potential purposes that it can be used for later. This results in a candid, unbiased content that is highly representative of an individual authoring it. The goal of *social filtering* is to identify a set of *"expert"* people whose posts are the most informative of prediction target and find an optimal way to transform them into a new feature space, which would yield the best performance for the model. The framework was evaluated on a case study of forecasting crime trend of Chicago, IL. We assessed different user representations, and individual's *positive affect* was proven to be the most informative for the application. We have also shown that the proper choice of an *expert "crowd"* is critical to prediction quality. The same pattern was also observed for individual crime categories. Experimental results suggest that user-generated content is rather predictive of future incidents than reflective on historical ones. It does not imply though that selected *experts* are manifesting their malicious intentions online, neither we claim the causality between tweets of average citizens and future incidents. Nonetheless, we have shown that individually aggregated content is able to capture latent factors that are predictive of prospective crimes.

Implications

This approach is not limited to crime trend prediction only and can potentially be applied to other tasks that support the notion of *user's quality*. In this case study we focused on a task where users were not explicitly sharing their opinions regarding

prediction target, this way guaranteeing absence of bias that is normally associated with publicly expressed beliefs. However, *social filtering* can be used in a setting where individuals are aware of the target[7], since it will automatically select the most diverse *experts*.

Future work

There are many directions for future research. Our results suggest that the model is highly dependent on transient topics discussed by selected users. While we were treating *positive affect* as a sufficient proxy, temporal topic modeling can be used to actually extract them. Also since we were interested to show feasibility of predicting trends by *social filtering* alone, in this work we only focused on *expert* individuals. However, incorporating socio-economic indexes, deemed to be correlates of criminal activity, can further enhance the quality of the model. Finer geographical resolution (*e.g.* districts, neighborhoods) as well as different locations can be examined. And finally, while we have shown that *social filtering* is able to select the most relevant people and extract hidden factors effective for such non-trivial task as crime forecasting, it would be interesting to see how it performs when applied to other domains.

[7]For instance, online platforms for investment communities solicit their users to express their insights regarding prospective market performance.

Table 4.1: Number of incidents reported per each of 29 categories

Type	Incidents
TOTAL	1,707,242
Theft	247,617
Battery	204,041
Narcotics	124,890
Criminal damage	120,934
Burglary	79,420
Assault	65,954
Other offense	63,676
Motor vehicle theft	57,227
Robbery	45,458
Deceptive practice	40,917
Criminal trespass	28,682
Weapons violation	12,408
Public peace violation	10,661
Offense involving children	7,343
Prostitution	7,311
Criminal sexual assault	4,330
Interference with public officer	3,840
Sex offense	3,344
Gambling	2,587
Liquor law violation	1,939
Homicide	1,547
Arson	1,542
Kidnapping	847
Stalking	581
Intimidation	524
Obscenity	96
Public indecency	43
Other narcotic violation	17
Non-criminal	14

Table 4.2: Statistics on user activity during period of observation. Total number of users: 2753. Total number of days: 1249. Users active more than average number of days: 1035 (83%); users with number of posts above average: 892 (32%).

	Days with posts	Overall tweets
min	1	1
avg	187	675
std	195	838
max	1237	3246

Chapter 5

Summary of Contributions and Future Work

5.1 Contributions

When it comes to applications relying on social media, the task of utmost importance is to ensure that the content used to base the insights on is credible, authentic and of high quality. The natural way to do so is within a context of an individual that authored it. With more studies resorting to a user-level aggregation in content sampling, it is necessary to address the challenges surrounding the approach and propose some feasible solutions.

In this work, we presented a *user-centric analytics* paradigm as prototype of a unified framework exploiting user streams for various purposes. By no means it should be considered completely generic and off-the-shelf solution, however, it was sufficient enough to shed some light on common issues that tend to appear when working with user timelines. Within its scope we studied three problems interesting and relevant on their own: topical experts detection, inference of absent opinions and

social filtering.

Detecting user's topical attribution is crucial for Twitter. While it is one of the most popular services for getting instant updates on professional matters, it still does not provide an explicit mechanism for community membership. Based on unique semantic signature of a group, we proposed a supervised approach to detecting relevance of a particular individual. Supervised models using static and dynamic representations of user-generated content were examined as well as a naïve lexicon-based filter. We also proposed a strategy for automatic generation of training data by exploiting class bias. Our experiments showed that while randomly sampling users from stream and assigning them to a negative class could potentially introduce some noise, it was robust enough to result in a high performance for both supervised models. Also neither *profile*- nor *behavior*-based models were dependent on a size of user's timeline, demonstrating good predictability equally for users with a small and extreme number of posts. Although naïve baseline somewhat underperformed, when compared to supervised models, it still yielded satisfactory results at a smaller cost. Thus, depending on requirements to an application using expert detection (*e.g.* real-time or not, tolerant to small ratio of false positives/negatives or not), either of appropriate models would work. We would like to point out that this component was only considered as a part of the holistic framework, and it was supposed to identify members of topical groups with a different level of expertise. Nonetheless, even in such setting it can be successfully used on its own for business intelligence purposes, automatic extraction of trends peculiar to specific domains and for professional follower recommendation.

Handling missing data becomes critical to applications depending on user streams. Either existing content was not captured due to platform rate limits, or a user actually was inactive during some time period—it results in tremendous amounts of time frames for which user's opinions cannot be retrieved. While this is one of the strong

arguments in support of bulk content, we deliberately chose user streams over the former because of the other advantages discussed earlier. We proposed a community-based approach to mitigate the effect of absent data. Missing opinions were modeled as a function of user's historical activity, opinions of his immediate network, conversations revolving in a community, and his propensity towards opinions of any kind. Expectedly, more sophisticated models performed better: *sentiment-based* with regards to a user-level assessment, and *content-based* on a micro level of individual opinions. The latter was also shown to be a viable initialization strategy for SVD matrix approximation, which as well can be applied to this problem. Our experimental results supported the hypothesis of high predictability dispersion of users from the same community. Interestingly, models of the most predictable individuals exhibited a high variance when trained on different data, which suggests that opinions of such users were timely addressing events that were triggering their conversations. We also presented a model examining user's content, activity and network characteristics to make a preliminary judgement on his predictability, which attained a satisfactory performance. We would like to note that while these models provide a reliable performance, one has to be cautious if using them solely to infer missing information. Even if the probability of interpolated data to be correct is extremely high, a user is not supposed to be liable for what is perceived as his "opinions". This problem has also implications for user privacy and if reversed can lead to a guidance on proper content protection.

Partially inspired by the concept of collaborative filtering, we presented *social filtering*—a user-centric framework for predicting social trends that operates on honest and unsolicited data. It works as a funnel that filters out content irrelevant to the problem and preserves only posts that are highly correlated with a target central to an outer application. We framed it as a three-step filter that selects active users relevant

to the problem and imposes an ensemble on it to refine the predictions. *Social filtering* attained a satisfactory performance even for a case study of crime trend prediction based on regular content of arbitrary citizens. The framework is generic enough to be utilized for any other application that considers user streams as a source of predictive signals.

5.2 Future Work

There are many ways in which this work can be continued. While the user-centric analytics is general enough, we focus here on each particular component.

Topical experts detection With all models showing decent performance, *behavior-based* is extremely expensive and not suitable for a real-time application, whereas test time of *profile-based* is relatively short and is attained by a small decrease in classification quality. Hence, it would be interesting to examine the performance of a hybrid model both in terms of prediction quality and resources consumption. Also we plan to investigate how the models should be adapted when applied to identification of broader groups (*e.g.* movie critics, social media experts, professional entertainers, *etc.*) compared to an example from a case study.

Interpolation of missing opinions In this study we examined fairly homogeneous community. It is interesting to see whether current models are capable of delivering similar results if applied to a broad group of people not united by a single motif (*e.g.* consider a person who uses Twitter only to maintain organic relationships with friends and chat on the topics of general interest). If it is not the case, we plan to study how the models should be adapted for such scenario. Also not all members of one's immediate community have a same impact on his decisions to publish content.

Thus the mechanism selecting individuals that are the most important to a target user is needed. While such data augmentation copes with missing opinions, it also results in added noise. We want to investigate to which extent opinion inference leads to improvement, and when the deterioration occurs. However, note that the latter would differ from context to context.

Social filtering While our goal in this work was to show that *social filtering* alone is able to extract predictive signals, it would be useful to augment such model with the knowledge pertaining to domain of a target application (in our case, we want to incorporate socio-economic indexes, historical weather records and other features characteristic of crime trend prediction). Also we would like to see how the model performs on the data for which users' opinions cannot be easily determined.

Bibliography

[1] Crime in the united states. Tech. rep., U.S. Department of Justice, FBI, 2013. Accessed: 2015-07-01.

[2] Population estimates, cities and towns. Tech. rep., U.S. Census Bureau, 2013. Accessed: 2014-12-20.

[3] Abel, F., Gao, Q., Houben, G., and Tao, K. Analyzing user modeling on twitter for personalized news recommendations. In *User Modeling, Adaption and Personalization - 19th International Conference, UMAP 2011, Girona, Spain, July 11-15, 2011. Proceedings* (2011), pp. 1–12.

[4] Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (2008), ACM, pp. 183–194.

[5] Althoff, T., and Leskovec, J. Donor retention in online crowdfunding communities: A case study of donorschoose.org. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015* (2015), pp. 34–44.

[6] Anderson, C. A. Heat and violence. *Current directions in psychological science 10*, 1 (2001), 33–38.

[7] Artzi, Y., Pantel, P., and Gamon, M. Predicting responses to microblog posts. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada* (2012), pp. 602–606.

[8] Balog, K., Azzopardi, L., and De Rijke, M. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (2006), ACM, pp. 43–50.

[9] Bar-Haim, R., Dinur, E., Feldman, R., Fresko, M., and Goldstein, G. Identifying and following expert investors in stock microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL* (2011), pp. 1310–1319.

[10] Baumer, E. P., and Wolff, K. T. Evaluating contemporary crime drop(s) in america, new york city, and many other places. *Justice Quarterly 31*, 1 (2014), 5–38.

[11] Becerra-Fernandez, I. Facilitating the online search of experts at nasa using expert seeker people-finder. In *PAKM* (2000), U. Reimer, Ed., vol. 34 of *CEUR Workshop Proceedings*, CEUR-WS.org.

[12] Bhattacharya, P., Ghosh, S., Kulshrestha, J., Mondal, M., Zafar, M. B., Ganguly, N., and Gummadi, K. P. Deep twitter diving: Exploring topical groups in microblogs at scale. In *Proceedings of the 17th ACM confer-*

ence on Computer supported cooperative work & social computing (2014), ACM, pp. 197–210.

[13] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *the Journal of machine Learning research 3* (2003), 993–1022.

[14] Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., and Pentland, A. Once upon a crime: Towards crime prediction from demographics and mobile data. *arXiv preprint arXiv:1409.2983* (2014).

[15] Bozzon, A., Brambilla, M., Ceri, S., Silvestri, M., and Vesci, G. Choosing the right crowd: expert finding in social networks. In *Proceedings of the 16th International Conference on Extending Database Technology* (2013), ACM, pp. 637–648.

[16] Braithwaite, J. *Crime, shame and reintegration*. Cambridge University Press, 1989.

[17] Cataldi, M., Di Caro, L., and Schifanella, C. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining* (2010), ACM, p. 4.

[18] Cetintas, S., Rogati, M., Si, L., and Fang, Y. Identifying similar people in professional social networks with discriminative probabilistic models. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (2011), ACM, pp. 1209–1210.

[19] Chainey, S., Tompson, L., and Uhlig, S. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal 21*, 1 (2008), 4–28.

[20] Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., and Leskovec, J. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web* (2014), ACM, pp. 925–936.

[21] Cheng, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. Antisocial behavior in online discussion communities. *arXiv preprint arXiv:1504.00680* (2015).

[22] Cheng, Z., Caverlee, J., Barthwal, H., and Bachani, V. Who is the barbecue king of texas?: a geo-spatial approach to finding local experts on twitter. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (2014), ACM, pp. 335–344.

[23] Chepurna, I., Aghababaei, S., and Makrehchi, M. How to predict social trends by mining user sentiments. In *2015 International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (2015), Springer, pp. 270–275.

[24] Chew, C., and Eysenbach, G. Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one 5*, 11 (2010), e14118.

[25] Cho, Y.-S., Galstyan, A., Brantingham, P. J., and Tita, G. Latent self-exciting point process model for spatial-temporal networks. *arXiv preprint arXiv:1302.2671* (2013).

[26] Christakis, N. A., and Fowler, J. H. The spread of obesity in a large social network over 32 years. *New England journal of medicine 357*, 4 (2007), 370–379.

[27] Collins, S., Sun, Y., Kosinski, M., Stillwell, D., and Markuzon, N. Are you satisfied with life?: Predicting satisfaction with life from facebook. In

2015 International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (2015), Springer, pp. 24–33.

[28] Crooks, A., Croitoru, A., Stefanidis, A., and Radzikowski, J. # earthquake: Twitter as a distributed sensor system. *Transactions in GIS 17*, 1 (2013), 124–147.

[29] Cullen, J. B., and Levitt, S. D. Crime, urban flight, and the consequences for cities. *Review of economics and statistics 81*, 2 (1999), 159–169.

[30] Culotta, A. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics* (2010), ACM, pp. 115–122.

[31] Darmon, D., Sylvester, J., Girvan, M., and Rand, W. Predictability of user behavior in social media: Bottom-up v. top-down modeling. *CoRR abs/1306.6111* (2013).

[32] Das, S., and Kramer, A. D. I. Self-censorship on facebook. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013.* (2013).

[33] De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. Predicting depression via social media. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013.* (2013).

[34] DeDeo, S. Collective phenomena and non-finite state computation in a human social system. *PloS one 8*, 10 (2013), e75818.

[35] Earle, P. S., Bowden, D. C., and Guy, M. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics 54*, 6 (2012).

[36] Eck, J., Chainey, S., Cameron, J., and Wilson, R. Mapping crime: Understanding hotspots. Tech. rep., U.S. Department of Justice, Office of Justice Programs, 2005.

[37] Ehrlich, I. On the relation between education and crime. *Education, income, and human behavior* (1975), 313–338.

[38] Fourney, A., White, R. W., and Horvitz, E. Exploring time-dependent concerns about pregnancy and childbirth from search logs. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), ACM, pp. 737–746.

[39] Freeman, R. B. The economics of crime. *Handbook of labor economics 3* (1999), 3529–3571.

[40] Gerber, M. S. Predicting crime using twitter and kernel density estimation. *Decision Support Systems 61* (2014), 115–125.

[41] Getoor, L., and Diehl, C. P. Link mining: a survey. *SIGKDD Explorations 7*, 2 (2005), 3–12.

[42] Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N., and Gummadi, K. Cognos: crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (2012), ACM, pp. 575–590.

[43] Gupta, M., Zhao, P., and Han, J. Evaluating event credibility on twitter. In *SDM* (2012), SIAM, pp. 153–164.

[44] Guy, I., Avraham, U., Carmel, D., Ur, S., Jacovi, M., and Ronen, I. Mining expertise and interests from social media. In *Proceedings of the 22nd international conference on World Wide Web* (2013), International World Wide Web Conferences Steering Committee, pp. 515–526.

[45] Hasan, M. A., and Zaki, M. J. A survey of link prediction in social networks. In *Social Network Data Analytics* (2011), pp. 243–275.

[46] Hill, S., and Ready-Campbell, N. Expert stock picker: the wisdom of (experts in) crowds. *International Journal of Electronic Commerce 15*, 3 (2011), 73–102.

[47] Hipp, J. R., Butts, C. T., Acton, R., Nagle, N. N., and Boessen, A. Extrapolative simulation of neighborhood networks based on population spatial distribution: Do they predict crime? *Social Networks 35*, 4 (2013), 614–625.

[48] Hogg, T., and Lerman, K. Social dynamics of digg. *CoRR abs/1202.0031* (2012).

[49] Hogg, T., Lerman, K., and Smith, L. M. Stochastic models predict user behavior in social media. *CoRR abs/1308.2705* (2013).

[50] Holcomb, J., Gottfried, J., Mitchell, A., and Schillinger, J. News use across social media platforms. Tech. rep., Pew Research Center, November 2013.

[51] Hong, Y., and Skiena, S. The wisdom of bookies? sentiment analysis versus. the nfl point spread. In *ICWSM* (2010).

[52] Joyce, E., and Kraut, R. E. Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication 11*, 3 (2006), 723–747.

[53] Kang, B., O'Donovan, J., and Höllerer, T. Modeling topic specific credibility on twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces* (2012), ACM, pp. 179–188.

[54] Kittur, A., Chi, E. H., and Suh, B. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2008), ACM, pp. 453–456.

[55] Lakkaraju, H., Leskovec, J., Kleinberg, J., and Mullainathan, S. A bayesian framework for modeling human evaluations.

[56] Lampos, V., De Bie, T., and Cristianini, N. Flu detector-tracking epidemics on twitter. In *Machine Learning and Knowledge Discovery in Databases*. Springer, 2010, pp. 599–602.

[57] Lauritsen, J. L., White, N., of Justice Statistics, B., of Justice, U. D., of Justice Programs, O., and of America, U. S. Seasonal patterns in criminal victimization trends.

[58] Lehmann, J., Castillo, C., Lalmas, M., and Zuckerman, E. Finding news curators in twitter. In *Proceedings of the 22nd international conference on World Wide Web companion* (2013), International World Wide Web Conferences Steering Committee, pp. 863–870.

[59] Li, J., and Cardie, C. Timeline generation: tracking individuals on twitter. In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014* (2014), pp. 643–652.

[60] Liao, W., Shah, S., and Makrehchi, M. Winning by following the winners: Mining the behaviour of stock market experts in social media. In *Social*

Computing, Behavioral-Cultural Modeling and Prediction - 7th International Conference, SBP 2014, Washington, DC, USA, April 1-4, 2014. Proceedings (2014), pp. 103–110.

[61] Liben-Nowell, D., and Kleinberg, J. M. The link-prediction problem for social networks. *Journal of the Association for Information Science and Technology 58*, 7 (2007), 1019–1031.

[62] Liu, J., Dolan, P., and Pedersen, E. R. Personalized news recommendation based on click behavior. In *Proceedings of the 2010 International Conference on Intelligent User Interfaces, February 7-10, 2010, Hong Kong, China* (2010), pp. 31–40.

[63] Livne, A., Simmons, M. P., Adar, E., and Adamic, L. A. The party is over here: Structure and content in the 2010 election. *ICWSM 11* (2011), 17–21.

[64] Llorente, A., Cebrian, M., Moro, E., et al. Social media fingerprints of unemployment. *arXiv preprint arXiv:1411.3140* (2014).

[65] Loughran, T., and McDonald, B. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance 66*, 1 (2011), 35–65.

[66] Lü, L., Zhang, Y.-C., Yeung, C. H., and Zhou, T. Leaders in social networks, the delicious case. *PloS one 6*, 6 (2011), e21202.

[67] Marwick, A. E., and Boyd, D. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society 13*, 1 (2011), 114–133.

[68] Mason, W., and Watts, D. J. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter 11*, 2 (2010), 100–108.

[69] Mathioudakis, M., and Koudas, N. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (2010), ACM, pp. 1155–1158.

[70] Mehlum, H., Moene, K., and Torvik, R. Crime induced poverty traps. *Journal of Development Economics 77*, 2 (2005), 325–340.

[71] Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. Self-exciting point process modeling of crime. *Journal of the American Statistical Association 106*, 493 (2011).

[72] Nikolov, S. *Trend or no trend: a novel nonparametric method for classifying time series*. PhD thesis, Massachusetts Institute of Technology, 2012.

[73] Pal, A., Chang, S., and Konstan, J. A. Evolution of experts in question answering communities. In *ICWSM* (2012).

[74] Pal, A., and Counts, S. Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining* (2011), ACM, pp. 45–54.

[75] Patterson, E. B. Poverty, income inequality, and community crime rates. *Criminology 29*, 4 (1991), 755–776.

[76] Paul, M. J., White, R. W., and Horvitz, E. Diagnoses, decisions, and outcomes: Web search as decision support for cancer. In *Proceedings of the 24th International Conference on World Wide Web* (2015), International World Wide Web Conferences Steering Committee, pp. 831–841.

[77] Pennebaker, J. W., Francis, M. E., and Booth, R. J. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates 71* (2001), 2001.

[78] Räbiger, S., and Spiliopoulou, M. A framework for validating the merit of properties that predict the influence of a twitter user. *Expert Systems with Applications 42*, 5 (2015), 2824–2834.

[79] Sakaki, T., Okazaki, M., and Matsuo, Y. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (2010), ACM, pp. 851–860.

[80] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., Park, G. J., Lakshmikanth, S. K., Jha, S., Seligman, M. E., et al. Characterizing geographic variation in well-being using tweets. In *ICWSM* (2013).

[81] Signorini, A., Segre, A. M., and Polgreen, P. M. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one 6*, 5 (2011), e19467.

[82] Sinha, S., Dyer, C., Gimpel, K., and Smith, N. A. Predicting the nfl using twitter. *arXiv preprint arXiv:1310.6998* (2013).

[83] Stankovic, M., Rowe, M., and Laublet, P. Finding co-solvers on twitter, with a little help from linked data. In *The Semantic Web: Research and Applications*. Springer, 2012, pp. 39–55.

[84] State, B., and Adamic, L. A. The diffusion of support in an online social movement: Evidence from the adoption of equal-sign profile pictures. In

Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14 - 18, 2015 (2015), pp. 1741–1750.

[85] Steeg, G. V., and Galstyan, A. Information transfer in social media. In *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012* (2012), pp. 509–518.

[86] Stomakhin, A., Short, M. B., and Bertozzi, A. L. Reconstruction of missing data in social networks based on temporal patterns of interactions. *Inverse Problems 27*, 11 (2011), 115013.

[87] Surowiecki, J. *The wisdom of crowds*. Anchor, 2005.

[88] Tita, G. E., and Boessen, A. Social networks and the ecology of crime: using social network data to understand the spatial distribution of crime. *The SAGE Handbook of Criminological Research Methods* (2011), 128.

[89] Traunmueller, M., Quattrone, G., and Capra, L. Mining mobile phone data to investigate urban crime theories at scale.

[90] Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM 10* (2010), 178–185.

[91] UzZaman, N., Blanco, R., and Matthews, M. Twitterpaul: Extracting and aggregating twitter predictions. *arXiv preprint arXiv:1211.6496* (2012).

[92] Valdes, J. M. D., Eisenstein, J., and De Choudhury, M. Psychological effects of urban crime gleaned from social media.

[93] Wagner, C., Liao, V., Pirolli, P., Nelson, L., and Strohmaier, M. It's not in their tweets: Modeling topical expertise of twitter users. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)* (2012), IEEE, pp. 91–100.

[94] Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M., Zheng, H., and Zhao, B. Y. Social turing tests: Crowdsourcing sybil detection. *arXiv preprint arXiv:1205.3856* (2012).

[95] Wang, G., Wang, T., Wang, B., Sambasivan, D., Zhang, Z., Zheng, H., and Zhao, B. Y. Crowds on wall street: Extracting value from collaborative investing platforms. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14 - 18, 2015* (2015), pp. 17–30.

[96] Wang, G. A., Jiao, J., Abrahams, A. S., Fan, W., and Zhang, Z. Expertrank: A topic-aware expert finding algorithm for online knowledge communities. *Decision Support Systems 54*, 3 (2013), 1442–1451.

[97] Wang, X., and Brown, D. E. The spatio-temporal modeling for criminal incidents. *Security Informatics 1*, 1 (2012), 1–17.

[98] Wang, X., Gerber, M. S., and Brown, D. E. Automatic crime prediction using events extracted from twitter posts. In *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer, 2012, pp. 231–238.

[99] Weng, J., Lim, E.-P., Jiang, J., and He, Q. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining* (2010), ACM, pp. 261–270.

[100] White, K., Li, G., and Japkowicz, N. Sampling online social networks using coupling from the past. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on* (2012), IEEE, pp. 266–272.

[101] Wilkinson, D. M. Strong regularities in online peer production. In *Proceedings 9th ACM Conference on Electronic Commerce (EC-2008), Chicago, IL, USA, June 8-12, 2008* (2008), pp. 302–309.

[102] Willnat, L., and Weaver, D. H. The american journalist in the digital age: Key findings. Tech. rep., School of Journalism, Indiana University, 2014.

[103] Xu, Z., Zhang, Y., Wu, Y., and Yang, Q. Modeling user posting behavior on social media. In *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012* (2012), pp. 545–554.

[104] Xue, Y., and Brown, D. E. Spatial analysis with preference specification of latent decision makers for criminal event prediction. *Decision support systems 41*, 3 (2006), 560–573.

[105] Yang, Y., and Pedersen, J. O. A comparative study on feature selection in text categorization. In *ICML* (1997), vol. 97, pp. 412–420.

[106] Yimam-Seid, D., and Kobsa, A. Expert-finding systems for organizations: Problem and domain analysis and the demoir approach. *Journal of Organizational Computing and Electronic Commerce 13*, 1 (2003), 1–24.

[107] Yin, H., Cui, B., Chen, L., Hu, Z., and Huang, Z. A temporal context-aware model for user behavior modeling in social media systems. In *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014* (2014), pp. 1543–1554.

[108] Yu, S., and Kak, S. A survey of prediction using social media. *arXiv preprint arXiv:1203.1647* (2012).

[109] Zafar, M. B., Bhattacharya, P., Ganguly, N., Gummadi, K. P., and Ghosh, S. Sampling content from online social networks: Comparing random vs. expert sampling of the twitter stream. *ACM Transactions on the Web (TWEB) 9*, 3 (2015), 12.

[110] Zhang, Y., Chen, W., Wang, D., and Yang, Q. User-click modeling for understanding and predicting search-behavior. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011* (2011), pp. 1388–1396.